Multiple sequence alignment

Multiple sequence alignment: today's goals

- to define what a multiple sequence alignment is and how it is generated; to describe profile HMMs
- to introduce databases of multiple sequence alignments
- to introduce ways you can make your own multiple sequence alignments
- to show how a multiple sequence alignment provides the basis for phylogenetic trees

Multiple sequence alignment: outline

[1] Introduction to MSA

- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

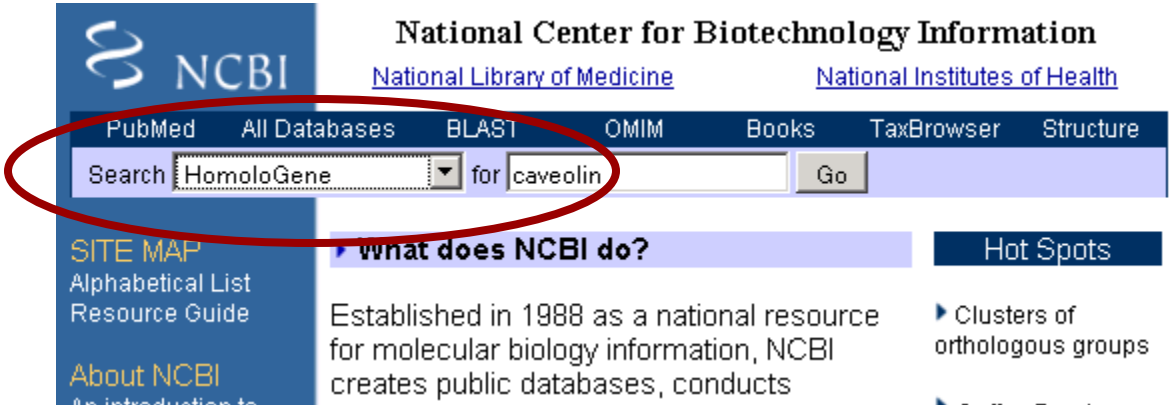
[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

Multiple sequence alignment: definition

- a collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned
- homologous residues are aligned in columns across the length of the sequences
- residues are homologous in an evolutionary sense
- residues are homologous in a structural sense

Example: someone is interested in caveolin



Step 1: at NCBI change the pulldown menu to HomoloGene and enter caveolin in the search box

Step 2: inspect the results. We'll take the first set of caveolins. Change the Display to Multiple alignment.

NCBI

HomoloGene
Discover Homologs

Help

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search HomoloGene for caveolin Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Send to

All: 4 | Fungi: 0 | Mammals: 0

Items 1 - 4 of 4

1: HomoloGene:1330. Gene conserved in Euteleostomi Download

CAV1	caveolin 1, caveolae protein, 22kDa	*Homo sapiens*
CAV1	caveolin 1, caveolae protein, 22kDa	*Pan troglodytes*
CAV1	caveolin 1, caveolae protein, 22kDa	*Canis lupus familiaris*
CAV1	caveolin 1, caveolae protein, 22kDa	*Bos taurus*
Cav1	caveolin 1, caveolae protein	*Mus musculus*
Cav1	caveolin 1, caveolae protein	*Rattus norvegicus*
CAV1	caveolin 1, caveolae protein, 22kDa	*Gallus gallus*
cav1	caveolin 1	*Danio rerio*

2: HomoloGene:7255. Gene conserved in Euteleostomi Download

CAV3	caveolin 3	*Homo sapiens*
CAV3	caveolin 3	*Pan troglodytes*
CAV3	caveolin 3	*Canis lupus familiaris*
CAV3	caveolin 3	*Bos taurus*
Cav3	caveolin 3	*Mus musculus*
Cav3	caveolin 3	*Rattus norvegicus*
CAV3	caveolin 3	*Gallus gallus*
cav3	caveolin 3	*Danio rerio*

Step 3: inspect the multiple alignment. Note that these eight proteins align nicely, although gaps must be included.

1: HomoloGene:1330. Gene conserved in Euteleostomi Downlo

Multiple Sequence Alignment

Generated by MUSCLE [see reference] versi 3.6 (using op on: -maxiters 2).

```
NP_001744.2          1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADELSEKQVYDAHT   46
XP_519325.2          1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADELSEKQVYDAHT   46
NP_001003296.1       1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-AEEMSEKQVYDAHT   46
NP_776429.1          1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-AEEMNEKQVYDAHT   46
NP_031642.1          1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADEVTEKQVYDAHT   46
NP_113744.1          1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADEVNEKQVYDAHT   46
XP_001234148.1       1   ---MEYFQ----EAFLYAAPVREQGNIYKPNNKMM-ADELSEKAVHDVHT   42
NP_997816.1          1   MTSG-YKDGTPEEEYAHSPFIRKQGNIYKPNNKEMDNDSINEKTLQDVHT   49

NP_001744.2         47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
XP_519325.2         47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
NP_001003296.1      47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
NP_776429.1         47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
NP_031642.1         47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
NP_113744.1         47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
XP_001234148.1      43   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   92
NP_997816.1         50   KEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTYSFDGVWKASFTTFTVTK   99
```

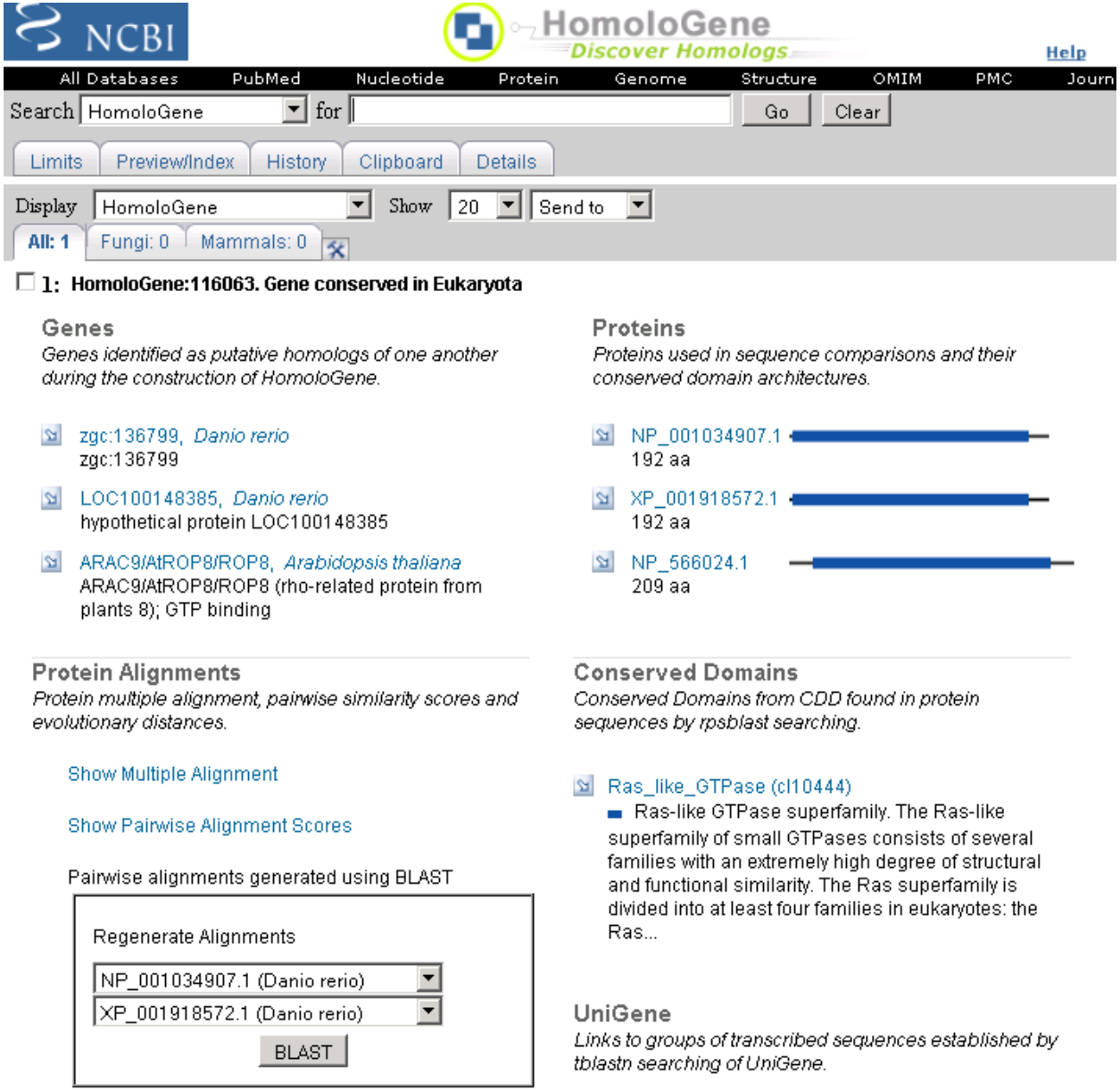
Here's another multiple alignment, Rac:

```
NP_061485.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
XP_855587.1          1   -----MQAIKCVVVEDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_776588.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_033033.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_599193.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_990348.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_956065.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_648121.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
XP_366655.1          1   MAAPGVQSLKCVVTGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSASVM   50
XP_329350.1          1   MLTGEMLTLDFLLL-------TCLLISYTTNAFPGEYIPTVFDNYSASVM   43
NP_195320.1          1   --MSASRFIKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV   48
NP_179371.1          1   --MSASRFIKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV   48
NP_190698.1          1   --MSASRFVKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV   48
NP_195228.1          1   --MSASRFIKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVI   48
NP_001048639.1       1   --MSASRFIKCVTVGDGAVGKTCMLISYTSNTFPTDYVPTVFDNFSANVV   48

NP_061485.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQTVGETYGKDITSRGKDKPIAD   95
XP_855587.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-------------------D   76
NP_776588.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-------------------D   76
NP_033033.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-------------------D   76
NP_599193.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-------------------D   76
NP_990348.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-------------------D   76
NP_956065.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-------------------D   76
NP_648121.1         46   VDAKPINLGLWDTAG
XP_366655.1         51   VDGKPISLGLWDTAG
XP_329350.1         44   VDGKPVSLGLWDTAG
NP_195320.1         49   VNGATVNLGLWDTAG
NP_179371.1         49   VNGATVNLGLWDTAG
NP_190698.1         49   VNGSTVNLGLWDTAGQEDYNRLRPLSYRGA-------------------D   79
NP_195228.1         49   VDGNTINLGLWDTAGQEDYNRLRPLSYRGA-------------------D   79
NP_001048639.1      49   VDGSTVNLGLWDTA--EDYNRLRPLSYRGA-------------------D   77
```

This insertion could be due to alternative splicing

HomoloGene includes groups of eukaryotic proteins. The site includes links to the proteins, pairwise alignments, and more



Example: 5 alignments of 5 globins

Let's look at a multiple sequence alignment (MSA) of five globins proteins. We'll use five prominent MSA programs: ClustalW, Praline, MUSCLE (used at HomoloGene), ProbCons, and TCoffee. Each program offers unique strengths.

We'll focus on a histidine (H) residue that has a critical role in binding oxygen in globins, and should be aligned. But often it's not aligned, and all five programs give different answers.

Our conclusion will be that there is no single best approach to MSA. Dozens of new programs have been introduced in recent years.

ClustalW

```
CLUSTAL W (1.83) multiple sequence alignment

                                                                    ▼
beta globin   ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin     -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin   -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean       ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
                            :   :   :   :  .. .    .      ::   *     *.
                            ▽                          ▼
beta globin   DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPE 102
myoglobin     HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLAS---LGRKHRAVGVKLS 104
soybean       --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice          --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
                .       . ..  *  .::        :           :              :

beta globin   NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
myoglobin     YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin   SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean       QFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA-------- 144
rice          HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
               :  :   ::  :        :      * .     .   :
```

Note how the region of a conserved histidine (▼) varies depending on which of five prominent algorithms is used

Praline

(a) Praline multiple sequence alignment

```
beta globin      ..........MVHLTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFES.FG
myoglobin        ...........MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK.FK
neuroglobin      .............MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean          ..........MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS..FL
rice             MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS..FL
Consistency      0000000000142654382579345734633643436244536864 33*35344*50063

beta globin      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSEL..HCDKLH....VDP
myoglobin        HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQS..HATKHK....IPV
neuroglobin      QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG....VKL
soybean          A.NGVDP..TNPKLTGHAEKLFALVRDSAGQL.KASGTVVADAA....LGSVHAQKAVTD
rice             R.NSDVPLEKNPKLKTHAMSVFVMTCEAAAQL.RKAGKVTVRDTTLKRLGATHLKYGVGD
Consistency      3166354224776653*436863542445445133563433354200333544 0000922

beta globin      ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKVVAGVANALAHKYH......
myoglobin        KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin      SSFSTVGESLLYMLEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE..
soybean          PQFVVVKEALLKTIKAAV.GDKWSDELSRAWEVAYDELAAAIKKA........
rice             AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE...
Consistency      437448444982585423053365544 54*5546542644675432200 1000
```

MUSCLE

(b)

```
MUSCLE (3.6) multiple sequence alignment

beta globin   ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin     -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin   -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean       ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
                             :   :   :   :  .. .    .      ::   *     *.

beta globin   DLSTPDAVMGNPKVKAHGKKVLGAF---SDGLAHLDNLKGTFATLSELHCDKLH--VDPE
myoglobin     HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAEIKPLAQSHATKHK--IPVK
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVI---DAAVTNVEDLSSLEEYLASLGRKHRAVGVKLS
soybean       NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice          NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
                             . ..  *  .::       :                  :    :

beta globin   NFRLLGNVLVCVLAHHFGKE-FTPPVQAAYQKVVAGVANALAHKYH------
myoglobin     YLEFISECIIQVLQSKHPGD-FGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin   SFSTVGESLLYMLEKCLGPA-FTPATRAAWSQLYGAVVQAMSRGWDGE----
soybean       QFVVVKEALLKTIKAAVGDK-WSDELSRAWEVAYDELAAAIKKA--------
rice          HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE---
               :  :   ::  :        :      * .     .   :
```

Probcons

(c)

```
PROBCONS

beta globin   M----------VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin     M-----------GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin   M-------------ERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean       M----------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
              *          * :   :   :   :  .. .    .      ::   *     *.

beta globin   DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP
myoglobin     HLKSEDEMKASEDLKKHGATVLTALGGI---LKKKGHHE---AEIKPLAQSHATKHKIPV
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLE---EYLASLGRKHRAV-GVKL
soybean       NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVV----ADAALGSVHAQK-AVTD
rice          NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKY-GVGD
              .    :    . ..  *  .::       ::   .           *.  *     :

beta globin   ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHK------YH
myoglobin     KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin   SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRG---W-DGE
soybean       PQFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIK--------KA
rice          AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
                :  :   ::  :        :      * .     .   :
```

TCoffee

(d)

```
CLUSTAL FORMAT for T-COFFEE Version_5.13

beta globin   ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
myoglobin     -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFK
neuroglobin   -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean       ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-FLA
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
                            :   :   :   :  .. .    .      ::   *     *.

beta globin   DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL---KGTF---ATLSELHCDKLHVDP
myoglobin     HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAE---IKPLAQSHATKHKIPV
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDL---SSLEEYLASLGRKH-RAVGVKL
soybean       NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice          NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
              .         . ..  *  .::        :               *.  *

beta globin   ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------
myoglobin     KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin   SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDG----E
soybean       Q-FVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA--------
rice          H-FEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
                :  :   ::  :        :      * .     .   :
```

Multiple sequence alignment: properties

- not necessarily one “correct” alignment of a protein family
- protein sequences evolve...
- ...the corresponding three-dimensional structures of proteins also evolve
- may be impossible to identify amino acid residues that align properly (structurally) throughout a multiple sequence alignment
- for two proteins sharing 30% amino acid identity, about 50% of the individual amino acids are superposable in the two structures

Multiple sequence alignment: features

- some aligned residues, such as cysteines that form disulfide bridges, may be highly conserved
- there may be conserved motifs such as a transmembrane domain
- there may be conserved secondary structure features
- there may be regions with consistent patterns of insertions or deletions (indels)

Multiple sequence alignment: uses

- MSA is more sensitive than pairwise alignment to detect homologs
- BLAST output can take the form of a MSA, and can reveal conserved residues or motifs
- Population data can be analyzed in a MSA (PopSet)
- A single query can be searched against a database of MSAs (e.g. PFAM)
- Regulatory regions of genes may have consensus sequences identifiable by MSA

Multiple sequence alignment: outline

[1] Introduction to MSA

- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

[5] Introduction to molecular evolution and phylogeny

Multiple sequence alignment: methods

Progressive methods: use a guide tree (related to a phylogenetic tree) to determine how to combine pairwise alignments one by one to create a multiple alignment.

Examples: CLUSTALW, MUSCLE

Multiple sequence alignment: methods

Example of MSA using ClustalW: two data sets

Five distantly related globins (human to plant)

Five closely related beta globins

Obtain your sequences in the FASTA format!
You can save them in a Word document or text editor.

Use ClustalW to do a progressive MSA

YOUR EM

KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def

MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

OUTPUT

OUTPUT FORMAT	OUTPUT ORDER
aln w/numbers	aligned

PHYLOGENETIC TREE

TREE TYPE	CORRECT DIST.	IGNORE GAPS
none	off	off

Enter or Paste a set of Sequences in any supported format:

Help

```
>beta_globin 2hhbB NP_000509.1 [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVL
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVA
ALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATV
TALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELF

. [Homo sapiens]
LFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVM
```

http://www.ebi.ac.uk/clustalw/

owse...

Run

Reset

Feng-Doolittle MSA occurs in 3 stages

[1] Do a set of global pairwise alignments (Needleman and Wunsch's dynamic programming algorithm)

[2] Create a guide tree

[3] Progressively align the sequences

Progressive MSA stage 1 of 3: generate global pairwise alignments

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	beta_globin	147	2	myoglobin	154	25
1	beta_globin	147	3	neuroglobin	151	15
1	beta_globin	147	4	soybean	144	13
1	beta_globin	147	5	rice	166	21
2	myoglobin	154	3	neuroglobin	151	16
2	myoglobin	154	4	soybean	144	8
2	myoglobin	154	5	rice	166	12
3	neuroglobin	151	4	soybean	144	17
3	neuroglobin	151	5	rice	166	18
4	soybean	144	5	rice	166	43

best score

Number of pairwise alignments needed

For n sequences, $(n-1)(n) / 2$

For 5 sequences, $(4)(5) / 2 = 10$

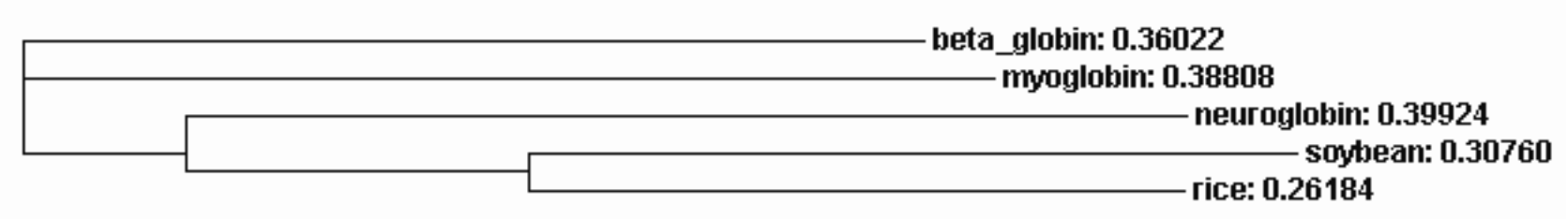
For 200 sequences, $(199)(200) / 2 = 19{,}900$

Feng-Doolittle stage 2: guide tree

- Convert similarity scores to distance scores
- A tree shows the distance between objects
- Use UPGMA (defined in the phylogeny lecture)
- ClustalW provides a syntax to describe the tree

Progressive MSA stage 2 of 3: generate a guide tree calculated from the distance matrix (5 distantly related globins)

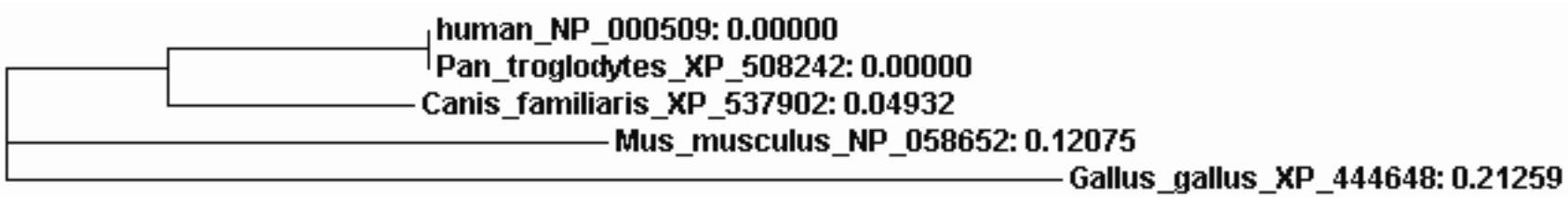
```
(
beta_globin:0.36022,
myoglobin:0.38808,
(
neuroglobin:0.39924,
(
soybean:0.30760,
rice:0.26184)
:0.13652)
:0.06560);
```



SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	human_NP_000509	147	2	Pan_troglodytes_XP_508242	147	100
1	human_NP_000509	147	3	Canis_familiaris_XP_537902	147	89
1	human_NP_000509	147	4	Mus_musculus_NP_058652	147	80
1	human_NP_000509	147	5	Gallus_gallus_XP_444648	147	69
2	Pan_troglodytes_XP_508242	147	3	Canis_familiaris_XP_537902	147	89
2	Pan_troglodytes_XP_508242	147	4	Mus_musculus_NP_058652	147	80
2	Pan_troglodytes_XP_508242	147	5	Gallus_gallus_XP_444648	147	69
3	Canis_familiaris_XP_537902	147	4	Mus_musculus_NP_058652	147	78
3	Canis_familiaris_XP_537902	147	5	Gallus_gallus_XP_444648	147	71
4	Mus_musculus_NP_058652	147	5	Gallus_gallus_XP_444648	147	66

```
(
(
(
human_NP_000509:0.00000,
Pan_troglodytes_XP_508242:0.00000)
:0.05272,
Canis_familiaris_XP_537902:0.04932)
:0.03231,
Mus_musculus_NP_058652:0.12075,
Gallus_gallus_XP_444648:0.21259);
```

5 closely related globins



Feng-Doolittle stage 3: progressive alignment

- Make a MSA based on the order in the guide tree
- Start with the two most closely related sequences
- Then add the next closest sequence
- Continue until all sequences are added to the MSA
- Rule: “once a gap, always a gap.”

Clustal W alignment of 5 distantly related globins

```
CLUSTAL W (1.83) multiple sequence alignment


beta_globin        ---------------MVHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFF 43
cytoglobin         MEKVPGEMEIERRERSEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFPSAKQYF 60
myoglobin          ----------------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKF 44
neuroglobin        ------------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLF 42
leghemoglobin      ----------------MGFTEKQEALVNSSWELFKQNPS-YSVLFYTIILKKAPAAKGMF 43
                                     :   :   :   *     .    .      ::   *     *

beta_globin        ES-FGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLA---HLDNLKGTFATLSELHCDKLHV 99
cytoglobin         SQ-FKHMEDPLEMERSPQLRKHACRVMGALNTVVENLHDPDKVSSVLALVGKAHALKHKV 119
myoglobin          DK-FKHLKSEDEMKASEDLKKHGATVLTALGGILK---KKGHHEAEIKPLAQSHATKHKI 100
neuroglobin        QYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG-V 101
leghemoglobin      S----FLKDSAEVVDSPKLQAHAEKVFGMVHDSAIQLRASGEVVLGDATLGAIHIQKGVV 99
                   .     :.       . ..  *   *:  .          .        :.  *     :

beta_globin        DPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------------ 147
cytoglobin         EPVYFKILSGVILEVVAEEFASDFPPETQRAWAKLRGLIYSHVTAAYKEVGWVQQVPNAT 179
myoglobin          PVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG------ 154
neuroglobin        KLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWDGE---------- 151
leghemoglobin      DP-HFVVVKEALLETIKEASGEKWSEELSTAWEVAYEGLASAIKKAMN------------ 146
                       :  :   ::  :       :      *       . . :    .

beta_globin        -----------
cytoglobin         TPPATLPSSGP 190
myoglobin          -----------
neuroglobin        -----------
leghemoglobin      -----------
```

Clustal W alignment of 5 closely related globins

```
CLUSTAL W (1.83) multiple sequence alignment


human_NP_000509                 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
Pan_troglodytes_XP_508242       MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
Canis_familiaris_XP_537902      MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLS 50
Mus_musculus_NP_058652          MVHLTDAEKSAVSCLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLS 50
Gallus_gallus_XP_444648         MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLS 50
                                *** *  **. :: **.***  * *.***.***:*******:* ***:**

human_NP_000509                 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100
Pan_troglodytes_XP_508242       TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100
Canis_familiaris_XP_537902      TPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVD 100
Mus_musculus_NP_058652          SASAIMGNPKVKAHGKKVITAFNEGLKNLDNLKGTFASLSELHCDKLHVD 100
Gallus_gallus_XP_444648         SPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVD 100
                                :. *::.*. *:******: :*.:.: :***:*.**: ************

human_NP_000509                 PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
Pan_troglodytes_XP_508242       PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
Canis_familiaris_XP_537902      PENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH 147
Mus_musculus_NP_058652          PENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH 147
Gallus_gallus_XP_444648         PENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH 147
                                ****:***: :: **. *:.*:***  ***:**:*  ** ***:***
```

* asterisks indicate identity in a column

Why “once a gap, always a gap”?

- There are many possible ways to make a MSA
- Where gaps are added is a critical question
- Gaps are often added to the first two (closest) sequences
- To change the initial gap choices later on would be to give more weight to distantly related sequences
- To maintain the initial gap choices is to trust that those gaps are most believable

Additional features of ClustalW improve its ability to generate accurate MSAs

- Individual weights are assigned to sequences; very closely related sequences are given less weight, while distantly related sequences are given more weight

- Scoring matrices are varied dependent on the presence of conserved or divergent sequences, e.g.:

PAM20	80-100% id
PAM60	60-80% id
PAM120	40-60% id
PAM350	0-40% id

- Residue-specific gap penalties are applied

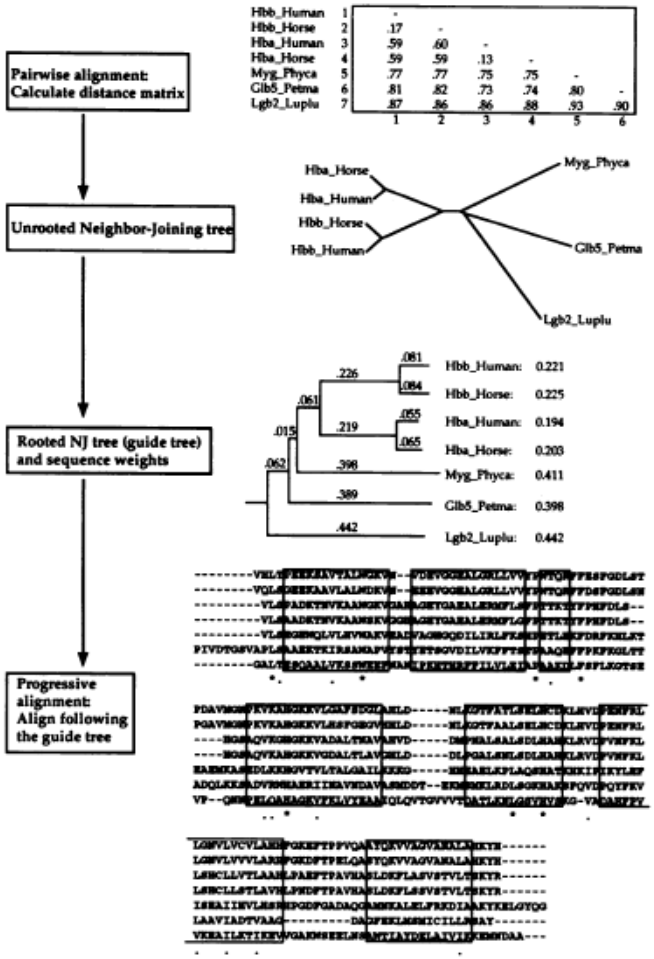


Figure 1. The basic progressive alignment procedure, illustrated using a set of 7 globins of known tertiary structure. The sequence names are from Swiss Prot (38): Hba__Horse: horse α-globin; Hba__Human: human α-globin; Hbb__Horse: horse β-globin; Hbb__Human: human β-globin; Myg__Phyca: sperm whale myoglobin; Glb5__Petma: lamprey cyanohaemoglobin; Lgb2__Luplu: lupin leghaemoglobin. In the distance matrix, the mean number of differences per residue is given. The unrooted tree shows all branch lengths drawn to scale. In the rooted

In Figure 1 we give the 7×7 distance matrix between the 7 globin sequences calculated using the full dynamic programming method.

The guide tree

The trees used to guide the final multiple alignment process are calculated from the distance matrix of step 1 using the Neighbour-Joining method (21). This produces unrooted trees with branch lengths proportional to estimated divergence along each branch. The root is placed by a 'mid-point' method (15) at a position where the means of the branch lengths on either side of the root are equal. These trees are also used to derive a weight for each sequence (15). The weights are dependent upon the distance from the root of the tree but sequences which have a common branch with other sequences share the weight derived from the shared branch. In the example in Figure 1, the leghaemoglobin (Lgb2__Luplu) gets a weight of 0.442, which is equal to the length of the branch from the root to it. The human β-globin (Hbb__Human) gets a weight consisting of the length of the branch leading to it that is not shared with any other sequences (0.081) plus half the length of the branch shared with the horse β-globin (0.226/2) plus one quarter the length of the branch shared by all four haemoglobins (0.061/4) plus one fifth the branch shared between the haemoglobins and myoglobin (0.015/5) plus one sixth the branch leading to all the vertebrate globins (0.062). This sums to a total of 0.221. In contrast, in the normal progressive alignment algorithm, all sequences would be equally weighted. The rooted tree with branch lengths and sequence weights for the 7 globins is given in Figure 1.

Progressive alignment

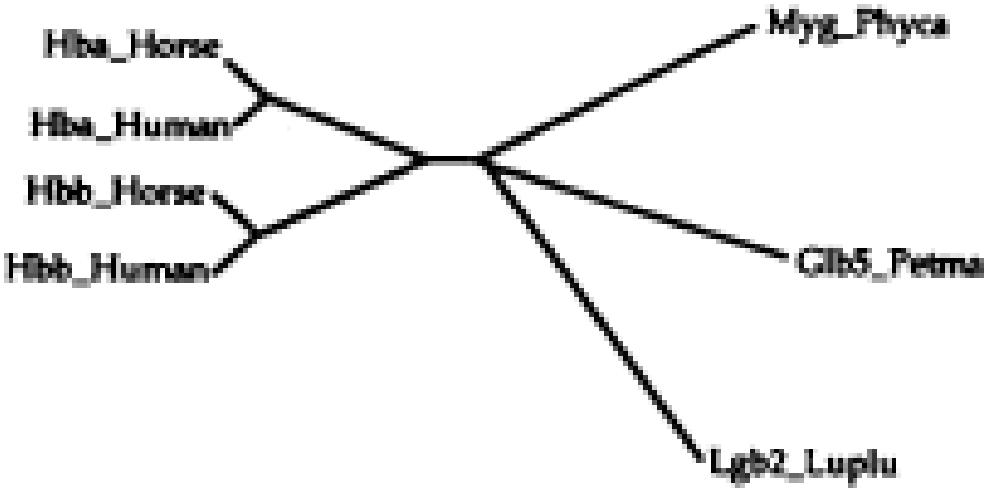
The basic procedure at this stage is to use a series of pairwise alignments to align larger and larger groups of sequences, following the branching order in the guide tree. You proceed from the tips of the rooted tree towards the root. In the globin example in Figure 1 you align the sequences in the following order: human vs. horse β-globin; human vs. horse α-globin; the 2 α-globins vs. the 2 β-globins; the myoglobin vs. the haemoglobins; the cyanohaemoglobin vs. the haemoglobins plus myoglobin; the leghaemoglobin vs. all the rest. At each stage

See Thompson et al. (1994) for an explanation of the three stages of progressive alignment implemented in ClustalW

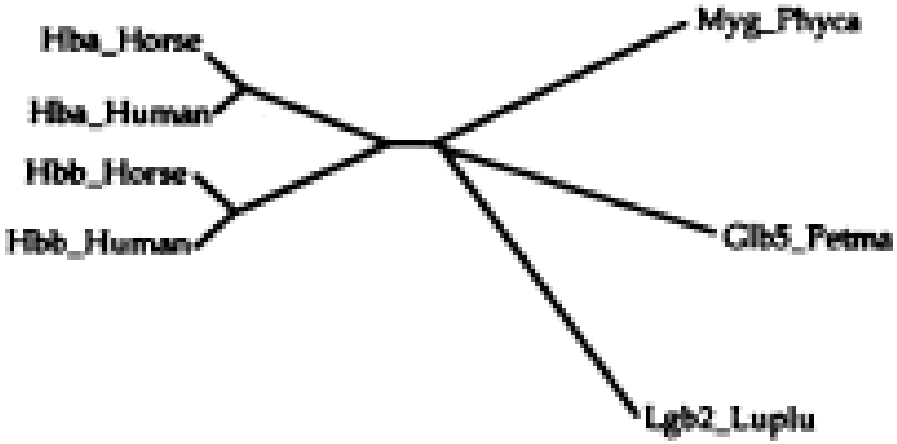
Pairwise alignment:
Calculate distance matrix

		1	2	3	4	5	6
Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90

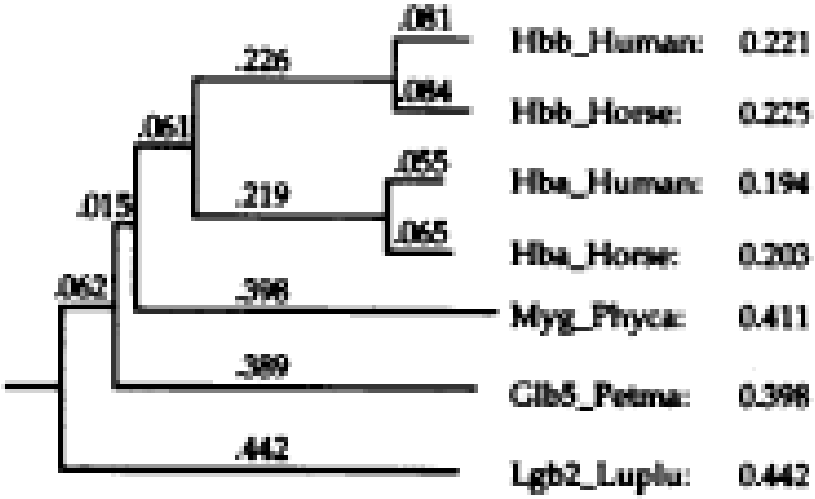
Unrooted neighbor-joining tree



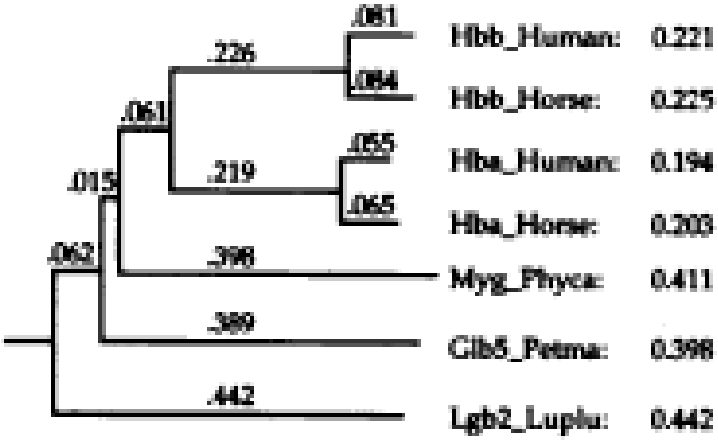
Unrooted neighbor-joining tree



Rooted neighbor-joining tree (guide tree) and sequence weights



Rooted neighbor-joining tree (guide tree) and sequence weights



Progressive alignment: Align following the guide tree

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

Multiple sequence alignment methods

Iterative methods: compute a sub-optimal solution and keep modifying that intelligently using dynamic programming or other methods until the solution converges.

Examples: MUSCLE, IterAlign, Praline, MAFFT

MUSCLE: next-generation progressive MSA

[1] Build a draft progressive alignment

- Determine pairwise similarity through k-mer counting (not by alignment)
- Compute distance (triangular distance) matrix
- Construct tree using UPGMA
- Construct draft progressive alignment following tree

MUSCLE: next-generation progressive MSA

[2] Improve the progressive alignment

- Compute pairwise identity through current MSA
- Construct new tree with Kimura distance measures
- Compare new and old trees: if improved, repeat this step, if not improved, then we're done

MUSCLE: next-generation progressive MSA

[3] Refinement of the MSA

- Split tree in half by deleting one edge
- Make profiles of each half of the tree
- Re-align the profiles
- Accept/reject the new alignment

Access to MUSLCE at EBI
http://www.ebi.ac.uk/muscle/

EMBL-EBI
European Bioinformatics Institute

Get | N

EBI Home | About EBI | Groups | Services | Toolbox | Databases | Downloads | Submissions

SEQUENCE ANALYSIS

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- Muscle Help
- Jalview Help
- Guide Tree
- Alignment
- Colours

- Similar Applications
  - ClustalW
  - T-Coffee

MUSCLE Submission Form

MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation.
MUSCLE is claimed to achieve both better average accuracy and better speed than CLUSTALW or T-Coffee, depending on the chosen options.

Download Software

MUSCLE
multiple sequence alignment method with reduced time and space complexity

EMAIL	RESULTS	ALIGNMENT TITLE	OUTPUT FORMAT	OUTPUT TREE
	interactive	Sequence	fasta	none

Enter or Paste a set of Sequences in any supported format: Help

Upload a file: Browse... Run Reset

Iterative approaches: MAFFT

- Uses Fast Fourier Transform to speed up profile alignment
- Uses fast two-stage method for building alignments using k-mer frequencies
- Offers many different scoring and aligning techniques
- One of the more accurate programs available
- Available as standalone or web interface
- Many output formats, including interactive phylogenetic trees

Iterative approaches: MAFFT

http://align.bmr.kyushu-u.ac.jp/mafft/online/server/

Grants ▾ NCBI Wikipedia Books etc ▾ Froogle Google Images SciCal Refworks Onc. Biostat Mail ▾ UCSC GB Expressionists GIANT BBC

EBI Tools: MUSCLE – MUltiple S... | MAFFT alignment and NJ / UPG... | Tree

MAFFT version 6

Multiple alignment program for amino acid or nucleotide sequences

- Download version
  - Mac OS X
  - Windows
  - Linux
  - Source
  - Usage
- **Online version**
  - **Alignment**
  - Phylogeny
- Merits and limitations
- Algorithms
- Tips
  - Aligning large data
  - Mafft-homologs
- Benchmarks
- Feedback

Contact address has changed!!

kkatoh@ kuicr.kyoto-u.ac.jp

katoh@ bioreg.kyushu-u.ac.jp

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:

Paste protein or DNA sequences in fasta format. Example

```
>gi|55743122|ref|NP_006735.2| retinol-binding protein 4, plasma precur
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL

>gi|12843160|dbj|BAB25881.1| unnamed protein product [Mus musculus]
MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDPEGLFLQDNIIAEFSVDEKGH
MSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRL
QNLDGTCADSYSFVFSRDPNGLSPETRRLVRQRQEELCLERQYRWIEHNGYCQSRPSRNSL

>gi|4502163|ref|NP_001638.1| apolipoprotein D precursor [Homo sapiens]
MVMLLLLLSALAGLFGAAEGQAFHLGKCPNPPVQENFDVNKYLGRWYEIEKIPTTFENGRCIQANYSLME
NGKIKVLNQELRADGTVNQIEGEATPVNLTEPAKLEVKFSWFMPSAPYWILATDYENYALVYSCTCIIQL
FHVDFAWILARNPNLPPETVDSLKNILTSNNIDVKKMTVTDQVNCPKLS
```

or upload a file: Browse...

Use structural alignment(s)

Output order:

- Same as input
- Aligned

Notify when finished (optional; recommended when submitting large data):

Email address:

Submit Reset

Advanced settings

Has about 1000 advanced settings!

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

Multiple sequence alignment: consistency

Consistency-based algorithms: generally use a database of both local high-scoring alignments and long-range global alignments to create a final alignment

These are very powerful, very fast, and very accurate methods

Examples: T-COFFEE, Prrp, DiAlign, ProbCons

ProbCons—consistency-based approach

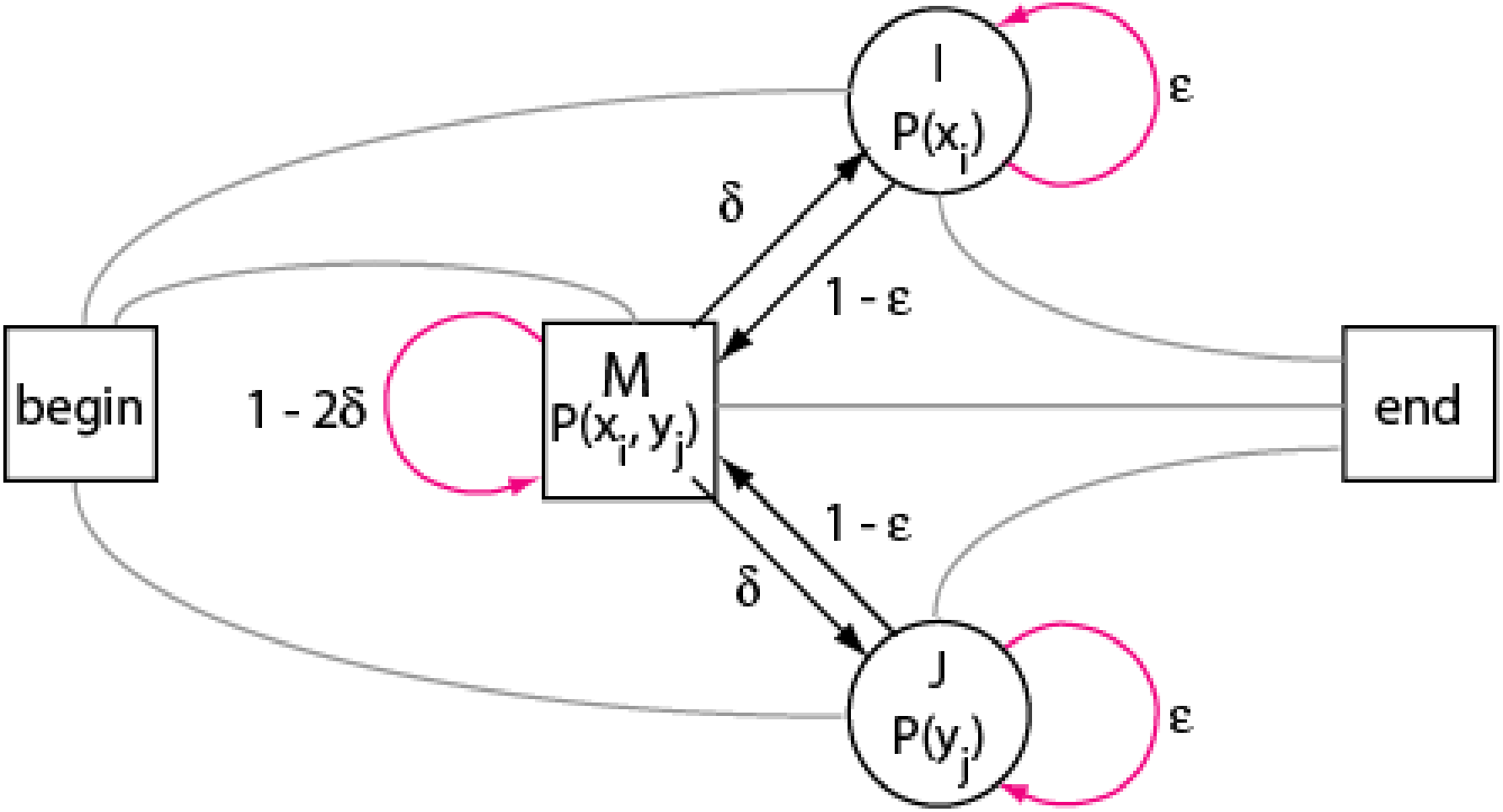
Combines iterative and progressive approaches with a unique probabilistic model.

Uses Hidden Markov Models to calculate probability matrices for matching residues, uses this to construct a guide tree

Progressive alignment hierarchically along guide tree

Post-processing and iterative refinement (a little like MUSCLE)

ProbCons uses an HMM to make alignments



ProbCons—consistency-based approach

Sequence x	x_i
Sequence y	y_j
Sequence z	z_k

If x_i aligns with z_k

and z_k aligns with y_j

then x_i should align with y_j

ProbCons incorporates evidence from multiple sequences to guide the creation of a pairwise alignment.

ProbCons output for the same alignment: consistency iteration helps

(c)

```
PROBCONS

beta globin   M----------VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin     M-----------GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin   M-------------ERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean       M----------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
              *          * :   :   :   :  .. .    .      ::   *     *.

beta globin   DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP
myoglobin     HLKSEDEMKASEDLKKHGATVLTALGGI---LKKKGHHE---AEIKPLAQSHATKHKIPV
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLE---EYLASLGRKHRAV-GVKL
soybean       NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVV----ADAALGSVHAQK-AVTD
rice          NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKY-GVGD
              .    :    . ..  *  .::        ::   .           *.  *      :

beta globin   ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHK------YH
myoglobin     KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin   SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRG---W-DGE
soybean       PQFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIK--------KA
rice          AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
                :  :   ::  :        :      * .     .   :
```

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

Access to TCoffee: http://tcoffee.org

Make a MSA
MSA w. structural data
Compare MSA methods
Make an RNA MSA
Combine MSA methods

Consistency-based
Structure-based

Back translate protein MSA

APDB ClustalW output:
TCoffee can incorporate structural information into a MSA

```
T-COFFEE, Version_4.71(Thu Nov 16 15:08:43 2006)
Cedric Notredame
CPU TIME:0 sec.
# APDB Evaluation: Color Range Blue-[0 % -- 100 %]-Red
# Sequence Score: APDB
# Local Score: APDB

SCORE=47
*
 BAD  AVG  GOOD
*
2hhbB    : 224
1V5HA    : 213
2MM1     : 219
1OJ6A    : 194
1FSL     : 157

2hhbB    ---------------MVHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYP
1V5HA   MEKVPGEMEIERRERSEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFP
2MM1     ----------------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHP
1OJ6A    ------------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEP
1FSL     ---------------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAP

                            :   :   :   :     .   .     ::   *
```

Protein Data Bank accession numbers

Multiple sequence alignment: outline

[1] Introduction to MSA

- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

Multiple sequence alignment: methods

How do we know which program to use?

There are benchmarking multiple alignment datasets that have been aligned painstakingly by hand, by structural similarity, or by extremely time- and memory-intensive automated exact algorithms.

Some programs have interfaces that are more user-friendly than others. And most programs are excellent so it depends on your preference.

If your proteins have 3D structures, **use these** to help you judge your alignments. For example, try Expresso at http://www.tcoffee.org.

Strategy for assessment of alternative multiple sequence alignment algorithms

[1] Create or obtain a database of protein sequences for which the 3D structure is known. Thus we can define “true” homologs using structural criteria.

[2] Try making multiple sequence alignments with many different sets of proteins (very related, very distant, few gaps, many gaps, insertions, outliers).

[3] Compare the answers.

BaliBase: comparison of multiple sequence alignment algorithms

```
Name                    hiv-1 protease

Number of sequences                 4
Alignment Length                    106
Longest Sequence                    104
Shortest Sequence                   98
Average Percent Identity            49
Maximum Percent Identity            86
Minimum Percent Identity            35

Sequence Name     SWISSPROT Accession
1fmb                  P32542
7upjB                 P03366
pol_sivcz             P17283
POL_SIVMK             P05897

Family   1fmb 7upjB pol_sivcz POL_SIVMK


1fmb          1  vTYNLEKRPTTIVLINDTPLNVLLDTGADTSVLTTahynrlkyrgrk.YQ
7upjB         1  pQFSLWKRPVVTAYIEGQPVEVLLDTGADDSIVAG.....iel.gnn.YS
pol_sivcz     1  pQITLWQRPLIPVKVEGQLCEALLDTGADDTVIER.....iqlqgl..WK
POL_SIVMK     1  pQFSLWRRPVVTAHIEGQPVEVLLDTGADDSIVTG.....iel.gph.YT

1fmb         50  GTGIGGVGGNVETFS.TPVTIKKKGRHIKTRMLVADIPVTILGRDILQDL
7upjB        44  PKIVGGIGGFINTLEYKNVEIEVLNKKVRATIMTGDTPINIFGRNILTAL
pol_sivcz    44  PKMIGGIGGFIKVKQFDNVHIEIEGRKVVGTVLVGPTPVNIIGRNILTQL
POL_SIVMK    44  PKIVGGIGGFINTKEYKNVEIEVLGKRIKRTIMTGDTPINIFGRNLLTAL

1fmb         99  GAKLVl
7upjB        94  GMSLNl
pol_sivcz    94  GCTLV.
POL_SIVMK    94  GMSLNl
```

Key

```
alpha helix         RED
beta strand         GREEN
core blocks         UNDERSCORE
```

Multiple sequence alignment: methods

Benchmarking tests suggest that ProbCons, a consistency-based/progressive algorithm, performs the best on the BAliBASE set, although MUSCLE, a progressive alignment package, is an extremely fast and accurate program.

ClustalW is the most popular program. It has a nice interface (especially with ClustalX) and is easy to use. But several programs perform better. There is no one single best program to use, and your answers will certainly differ (especially if you align divergent protein or DNA sequences)

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

Multiple sequence alignment to profile HMMs

► Hidden Markov models (HMMs) are “states” that describe the probability of having a particular amino acid residue at arranged in a column of a multiple sequence alignment

► HMMs are probabilistic models

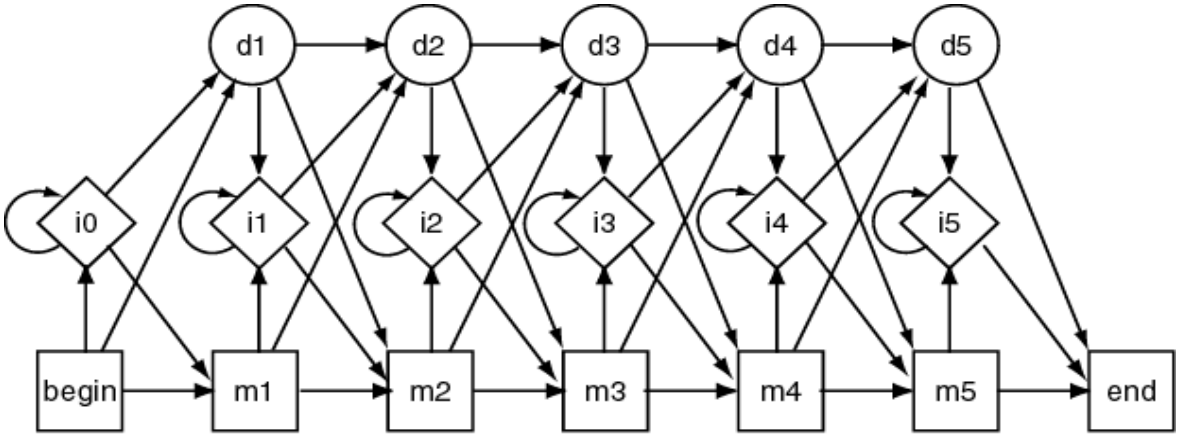
► HMMs may give more sensitive alignments than traditional techniques such as progressive alignment

Structure of a hidden Markov model (HMM)

delete state

insert state

main state



PFAM (protein family) database is a leading resource for the analysis of protein families

http://pfam.sanger.ac.uk/

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Pfam 24.0 (October 2009, 11912 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID Go Example

PFAM HMM for lipocalins: resembles a position-specific scoring matrix

```
HMMER2.0  [2.2f]
NAME  lipocalin
ACC   PF00061
DESC  Lipocalin / cytosolic fatty-acid binding protein family
LENG  157
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   hmmbuild -f -F HMM.ann SEED.ann
COM   hmmcalibrate --seed 0 HMM.ann
NSEQ  58
DATE  Tue May 29 15:09:42 2001
CKSUM 9538
GA    9.9 9.9
TC    9.9 9.9
NC    9.8 9.8
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT      -4  -8455
NULE     595  -1558     85    338   -294    453  -1158    197    249    902  -1085   -142    -21   -313     45    531    201    384  -1998   -644
EVD    -9.792953   0.668810
HMM        A      C      D      E      F      G      H      I      K      L      M      N      P      Q      R      S      T      V      W      Y
         m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
          -30      *  -5585
    1    1001  -4300    293    782  -4621  -3800  -2458  -4371   2047  -4316  -3389   -142  -1206   1489   1167     85  -2765  -3922   1025  -3800
    -    -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
    -      -6 -10723 -11765   -894  -1115   -701  -1378  -1030  -8285
    2   -3118  -2943  -5463  -4828   3095  -4667  -3535    901  -4424    483  -2138  -4313  -4716  -1252  -4224  -1588  -1107   1271  -3396   2265
    -    -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
    -      -6 -10723 -11765   -894  -1115   -701  -1378  -8316  -8281
    3    1590  -4263  -1018      1  -4568  -3808    -28  -4307  -2055    793  -3356    501  -3901     46  -1145    986    707    335  -4456   -655
    -    -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
                     -11765   -894  -1115   -701  -1378  -8316  -8276
                       -732  -4024  -7178   3700   -383  -7147  -5560  -7210  -6440  -4176  -5498  -4842  -6218  -1624  -4691  -1807  -7337  -6645
                        233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
                     -11765   -894  -1115   -701  -1378  -8316  -8271
```

20 amino acids

position

PFAM HMM for lipocalins: GXW motif

20 amino acids

22	528	-4297	985	1876	-4618	-3799	-2458	-1465	1726	-849	-3387	-1158	-3892	2	745	249	-2764	-1469	-4481	-3798
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-6	-10723	-11765	-894	-1115	-701	-1378	-8316	-8185											
23	-757	2219	-4989	-936	-1018	166	-607	391	550	929	1276	1400	-4623	-3765	-4013	-267	-105	37	-3424	-444
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-6	-10723	-11765	-894	-1115	-701	-1378	-8316	-8180											
24	577	-4296	922	-391	-4615	2193	945	474	-2039	-4311	-3385	-995	-1139	1027	-163	-838	-1123	-855	-4479	-3797
-	-149	-500	233	43	-381	[illegible]	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-6	-10723	-11765	-894	-1115	-701	-1378	-8316	-8175											
25	-372	260	-2807	-1166	-4301	-1825	-253	-686	129	431	-3189	-185	1682	699	332	168	-1466	1338	573	-3696
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-6	-10723	-11765	-894	-1115	-701	-1378	-8316	-8170											
26	861	-4204	78	-393	349	-465	-2489	-203	-842	774	1840	711	285	-724	-953	185	-684	-1776	-4413	-3757
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-6	-10723	-11765	-894	-1115	-701	-1378	-8316	-8165											
27	-800	-4296	449	1532	735	-3799	58	-132	366	-1776	-3385	-167	-1095	941	1752	-63	-1120	-1835	306	-3798
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-6	-10723	-11765	-894	-1115	-701	-1378	-8316	-8160											
28	1682	-2978	-206	-400	593	-4564	-3409	-990	-1311	-1859	2388	-4052	-4619	-3749	-4002	-670	351	1354	-3427	1575
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-6	-10723	-11765	-894	-1115	-701	-1378	-8316	-8155											

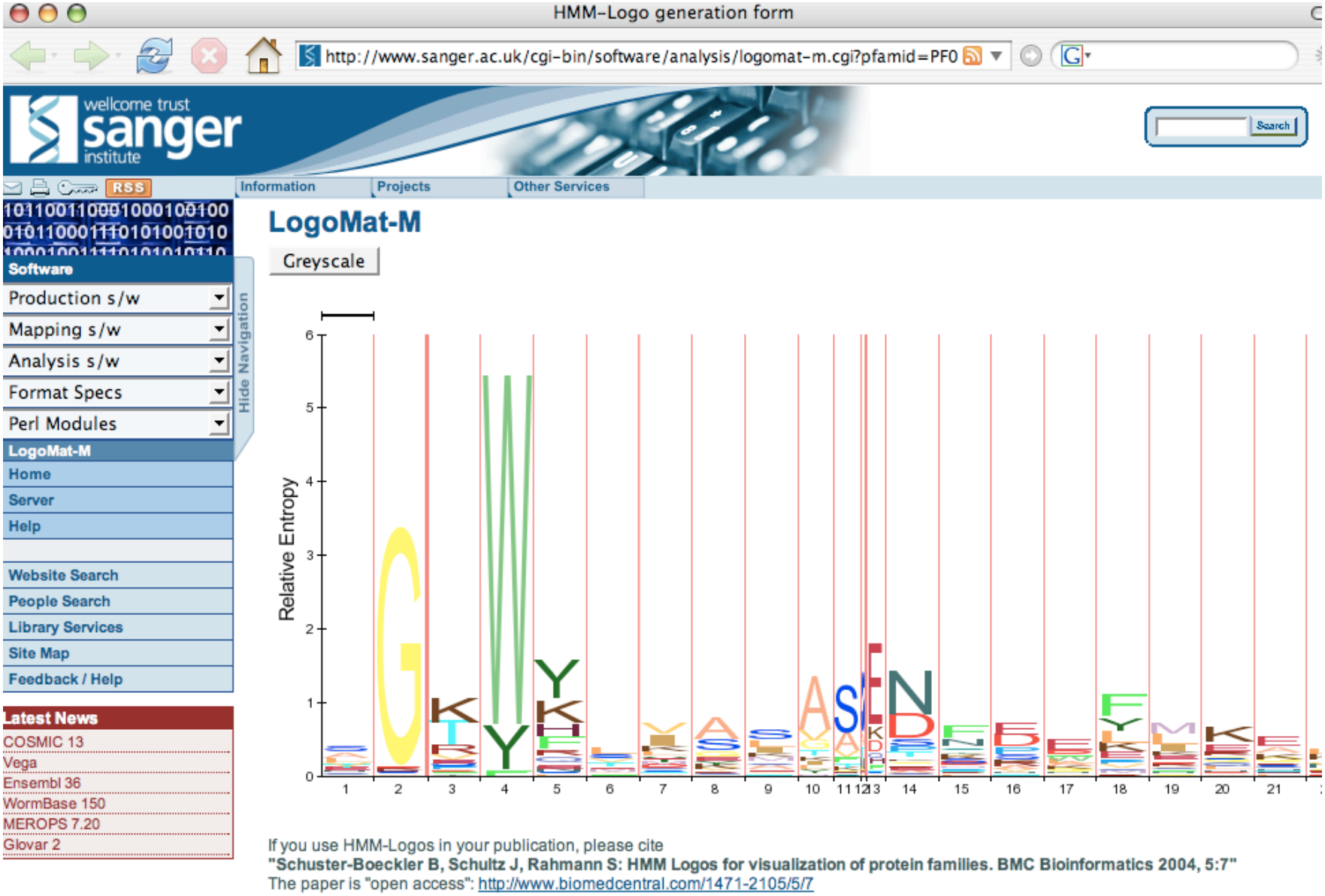
G

W

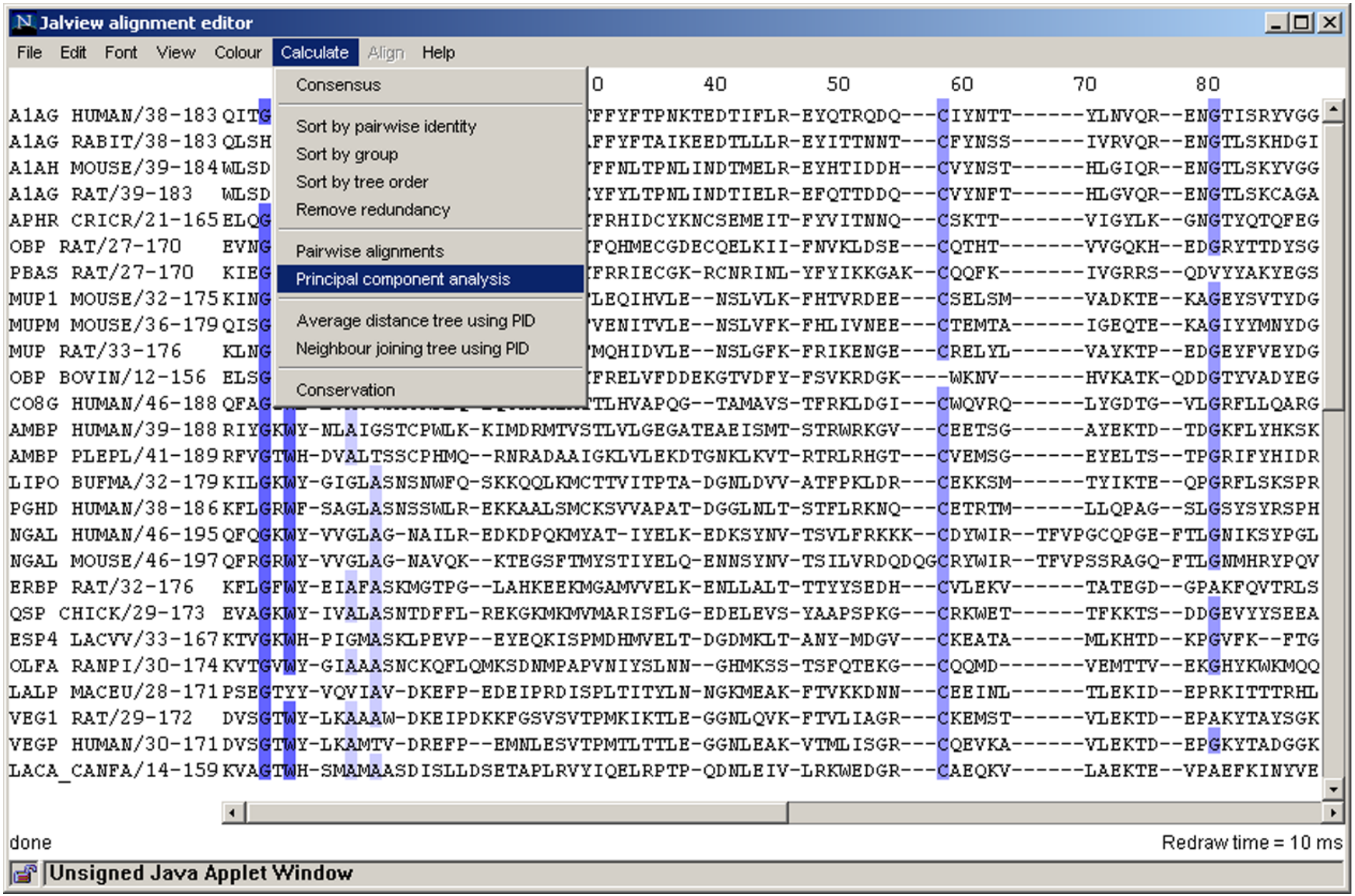
PFAM GCG MSF format

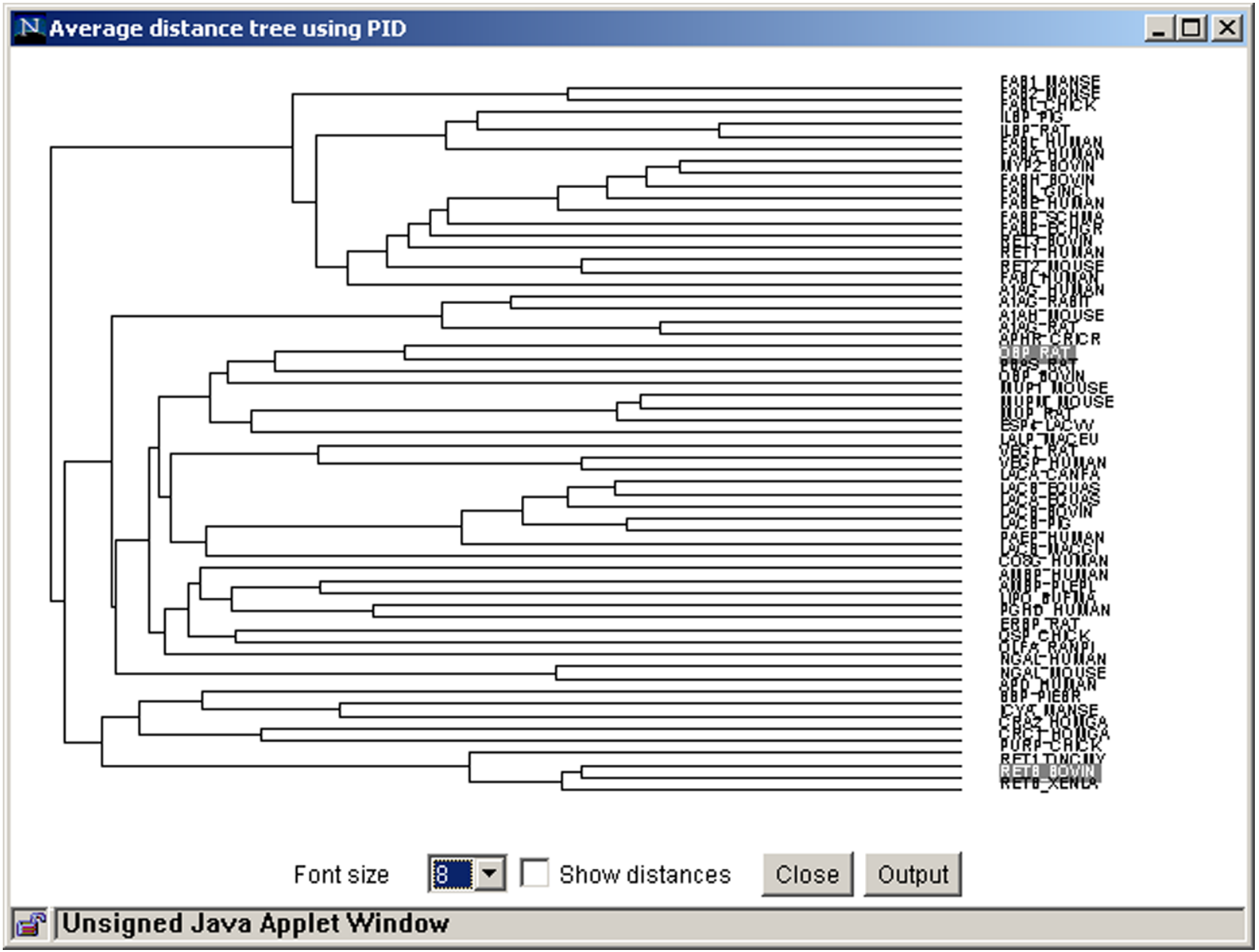
```
                        1                                                  50
A1AG_HUMAN_38-183       QITGKWF.YI ASAFRNEEYN .KSVQEIQAT FFYFTPNKTE DTIFLR.EYQ
A1AG_RABIT_38-183       QLSHKWF.FT ASAFRNPKYK .QLVQHTQAA FFYFTAIKEE DTLLLR.EYI
A1AH_MOUSE_39-184       WLSDKWF.FI GAAVLNPDYR .QEIQKTQMV FFNLTPNLIN DTMELR.EYH
A1AG_RAT_39-183         WLSDKWF.YM GAAFRDPVFK .QAVQTIQTE YFYLTPNLIN DTIELR.EFQ
APHR_CRICR_21-165       ELQGKWY.TI VIAADNLEKI .EEGGPLRFY FRHIDCYKNC SEMEIT.FYV
OBP_RAT_27-170          EVNGDWR.TL YIVADNVEKV .AEGGSLRAY FQHMECGDEC QELKII.FNV
PBAS_RAT_27-170         KIEGNWR.TV YLAASSVEKI .NEGSPLRTY FRRIECGK.R CNRINL.YFY
MUP1_MOUSE_32-175       KINGEWH.TI ILASDKREKI .EDNGNFRLF LEQIHVLE.. NSLVLK.FHT
MUPM_MOUSE_36-179       QISGYWF.SI AEASYEREKI .EEHGSMRAF VENITVLE.. NSLVFK.FHL
MUP_RAT_33-176          KLNGDWF.SI VVASNKREKI .EENGSMRVF MQHIDVLE.. NSLGFK.FRI
OBP_BOVIN_12-156        ELSGPWR.TV YIGSTNPEKI .QENGPFRTY FRELVFDDEK GTVDFY.FSV
CO8G_HUMAN_46-188       QFAGTWL.LV AVGSACRFLQ .EQGHRAEAT TLHVAPQG.. TAMAVS.TFR
AMBP_HUMAN_39-188       RIYGKWY.NL AIGSTCPWLK .KIMDRMTVS TLVLGEGATE AEISMT.STR
AMBP_PLEPL_41-189       RFVGTWH.DV ALTSSCPHMQ ..RNRADAAI GKLVLEKDTG NKLKVT.RTR
LIPO_BUFMA_32-179       KILGKWY.GI GLASNSNWFQ .SKKQQLKMC TTVITPTA.D GNLDVV.ATF
PGHD_HUMAN_38-186       KFLGRWF.SA GLASNSSWLR .EKKAALSMC KSVVAPAT.D GGLNLT.STF
NGAL_HUMAN_46-195       QFQGKWY.VV GLAG.NAILR .EDKDPQKMY AT.IYELK.E DKSYNV.TSV
NGAL_MOUSE_46-197       QFRGRWY.VV GLAG.NAVQK ..KTEGSFTM YSTIYELQ.E NNSYNV.TSI
ERBP_RAT_32-176         KFLGFWY.EI AFASKMGTPG ..LAHKEEKM GAMVVELK.E NLLALT.TTY
QSP_CHICK_29-173        EVAGKWY.IV ALASNTDFFL .REKGKMKMV MARISFLG.E DELEVS.YAA
ESP4_LACVV_33-167       KTVGKWH.PI GMASKLPEVP ..EYEQKISP MDHMVELT.D GDMKLT.ANY
OLFA_RANPI_30-174       KVTGVWY.GI AAASNCKQFL QMKSDNMPAP VNIYSLNN.. GHMKSS.TSF
LALP_MACEU_28-171       PSEGTYY.VQ VIAV.DKEFP .EDEIPRDIS PLTITYLN.N GKMEAK.FTV
VEG1_RAT_29-172         DVSGTWY.LK AAAW.DKEIP DKKFGSVSVT PMKIKTLE.G GNLQVK.FTV
```

Pfam (protein family) database



PFAM JalView viewer




Average distance tree using PID
FAB1_MANSE
FAB2_MANSE
FABL_CHICK
LBP_PIG
LBP_RAT
FABE_HUMAN
FABA_HUMAN
MYP2_BOVIN
FABH_BOVIN
FAB1_GINGI
FABP_HUMAN
FABP_SCHMA
FABP_BCHGR
RET3_BOVIN
RET1_HUMAN
RET2_MOUSE
FABC_HUMAN
A1AG_HUMAN
A1AG_RABIT
A1AH_MOUSE
A1AG_RAT
APHR_CRICR
OBP_RAT
PROS_RAT
OBP_BOVIN
MUP1_MOUSE
MUP4_MOUSE
MUP_RAT
ESP4_LACVV
LALP_MACEU
VEST_RAT
VESP_HUMAN
LACA_CANFA
LACB_EQUAS
LACA_EQUAS
LACB_BOVIN
LACB_PIG
PAEP_HUMAN
LACB_MACGI
CO8G_HUMAN
AMBP_HUMAN
AMBP_PLEBI
LIPO_BUFMA
PGRD_HUMAN
EBBP_RAT
OSP_CHICK
OLFA_RANPI
NGAL_HUMAN
NGAL_MOUSE
APOD_HUMAN
BBP_PIEBR
CYS_MANSE
CRB2_HOMGA
CRC1_HOMGA
PURP_CHICK
RET1_TINGUY
RET8_BOVIN
RET8_XENLA
Font size
8
Show distances
Close
Output
Unsigned Java Applet Window

Sequence analysis

You may use either the swissprot/sptrembl sequence identifier (ID) / accession number (ACC) or the protein sequence itself to request the smart service

Sequence ID or ACC

Sequence

```
>gi|5803139|ref|NP_006735.1| retinol-binding
protein 4, plasma precursor; retinol-binding
protein 4, plasma [Homo sapiens]
MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMA
KKDPEGLFLQDNIVAEFSVDETGQMS
ATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQ
KGNDDHWIVDTDYDTYAVQYSCRLLN
LDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLI
```

Sequence SMART | Reset

HMMER searches of the SMART database occur by default. You may also find:

- [x] Outlier homologues and homologues of known structure
- [ ] PFAM domains
- [ ] signal peptides
- [ ] internal repeats

Architecture analysis

You can search for proteins with combinations of specific domains in different species or taxonomic ranges.

Domain selection

Examples: TyrKc AND SH3 AND NOT SH2
UNIQUE SH2

Taxonomic selection

Select a taxonomic range via the selection box or type it into the text box below:

All

Examples: Dictyostelium discoideum
Porifera

Architecture SMART | Reset

You can try an Advanced Query if you're familiar with SQL.

Alert

If you want to be automatically informed each time a new protein with a defined domain composition is deposited in databases, please use 'alert SMART' (this facility is also available following an architecture analysis query)

Domains detected by SMART

You can search for keywords in the domain annotation

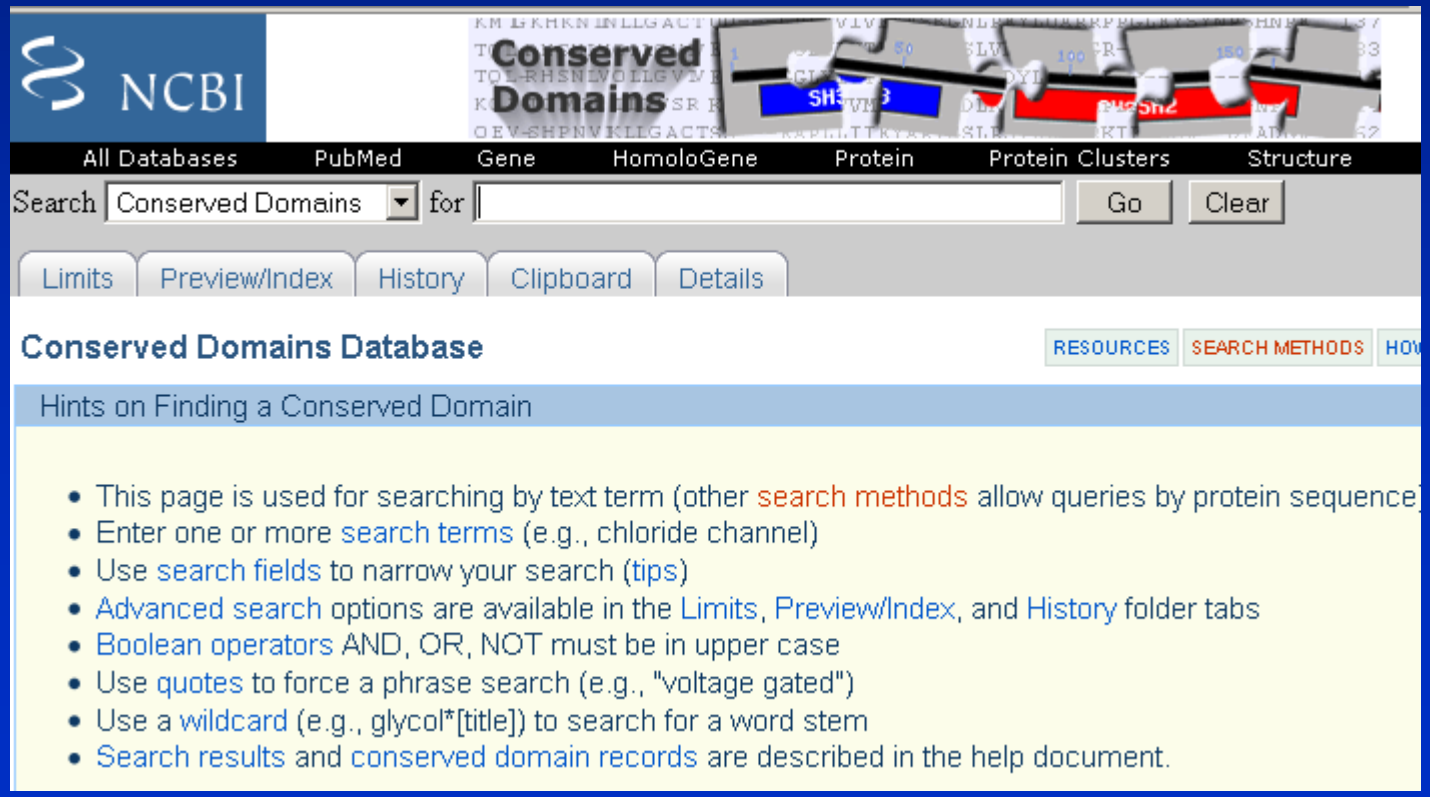
Search Keywords

- Browse the database of all available domains in the SMART databas
- See a list of recent domain changes

SMART: Simple Modular Architecture Research Tool (emphasis on cell signaling)

CDD: Conserved domain database (at NCBI): **CDD = Pfam + SMART**

[1] Go to NCBI → Domains & Structure (left sidebar)
[2] Click CDD
[3] Enter a text query, or a protein sequence



CDD entry for “globin”

NCBI

Conserved Domains

All Databases | PubMed | Gene | HomoloGene | Protein | Protein Clusters | Structure

Search Conserved Domains for globin Go Clear Save S

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort By Send to

Links: Related CDs, Literature, Sequence, Structure, BioSystems, Other Links

All: 9 | NCBI-curated: 3 | families: 7 | imported: 4 | superfamilies: 4

Items 1 - 3 of 3 One page.

Recent acti

1: cd01040 Related CDs, Literature, Sequence, Structure, BioSystems, Other Links

globin:
Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen in the bloodstream, (2) microorganismal flavohemoglobins, which are linked to C-terminal FAD-dependend reductase domains, (3) homodimeric bacterial hemoglobins, such as from Vitreoscilla, (4) plant leghemoglobins (symbiotic hemoglobins, involved in nitrogen metabolism in plant rhizomes), (5) plant non-symbiotic hexacoordinate globins and hexacoordinate globins from bacteria and animals, such as neuroglobin, (6) invertebrate hemoglobins, which may occur in tandem-repeat arrangements, and (7) monomeric myoglobins found in animal muscle tissue. [29979]

CDD = PFAM + SMART

CDD entry for “globin”

NCBI

Conserved Domains

SH3 SH2

HOME | SEARCH | SITE MAP | Entrez | CDD | Structure | Protein | Help

cd01040: globin

Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen in the bloodstream, (2) microorganismal flavohemoglobins, which are linked to C-terminal FAD-dependend reductase domains, (3) homodimeric bacterial hemoglobins, such as from Vitreoscilla, (4) plant leghemoglobins (symbiotic hemoglobins, involved in nitrogen metabolism in plant rhizomes), (5) plant non-symbiotic hexacoordinate globins and hexacoordinate globins from bacteria and animals, such as neuroglobin, (6) invertebrate hemoglobins, which may occur in tandem-repeat arrangements, and (7) monomeric myoglobins found in animal muscle tissue.

Links

Source: Cdd
Taxonomy: cellular organisms
PubMed: 5 links
Book: 3 links
Protein: Representatives
Specific Protein
Related Protein
Related Structure
Architectures
Superfamily: cl00280
BioSystems: 19 links

Statistics

PSSM-Id: 29979
View PSSM: cd01040
Aligned: 203 rows
Threshold Bit Score: 50.4265
Threshold Setting Gi: 15599162
Created: 10-Jan-2006
Updated: 10-Jan-2006

Structure

Structure View
Program: Cn3D
Drawing: All Atoms
Aligned Rows: up to 10
Download Cn3D

Hierarchy

Interactive Display
Display: cd01040 Branch

Conserved Features/Sites | PubMed References | Book References

heme-binding

Feature 1: heme-binding site

Evidence:
- **Structure:** Ascaris hemoglobin with bound heme and oxygen molecule
  - View structure with Cn3D
- **Comment:** Ascaris hemoglobin exhibits strong affinity to oxygen
- **Citation:** PMID 7753786
- **Structure:** Bovine deoxy-hemoglobin A with bound heme
  - View structure with Cn3D
- **Citation:** PMID 8411160

Download Cn3D for Viewing 3D Structure | Scroll to Sequence Alignment Display

cd01040 is part of a hierarchy of related CD models.
Use the graphical representation to navigate this hierarchy.
cd01040 is a member of the superfamily cl00280.

cd01040 Sequence Cluster

Zoom In | Detailed View

Sub-family Hierarchy

Interactive Display with CDTree

- cd01067 globin_like
  - cd00454 truncated_globin_I
  - cd01040 globin
  - cd01068 sensor_globin

CDD entry for “globin”

Sequence Alignment

Reformat | Format: Hypertext | Row Display: up to 10 | Color Bits: 2.0 bit | Type Selection: the most diverse members

```
                          10        20        30        40        50        60        70        80
                  ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
Feature 1                                   #                   #  ##                      #  #
1ASH          1   ANKTRELCMKSLehakvd--tsnearQDGIDLYKHMFENYp------pLRKYFksreeytaedvqndpfFAKQGQKILLA  72
1HDA_A        3   SAADKGNVKAAWgkvg------ghaaEYGAEALERMFLSFp------tTKTYFphfdls-----hgsaqVKGHGAKVAAA  65
gi 422406     6   SDSEEKLVRDAWapih------gdlqGTANTVFYNYLKKYp------sNQDKFetlkghpldevkdtanFKLIAGRIFTI  73
gi 1730834  167   TARPRKTKQRDNdnkv-------dtaLFCSQFYDNLIAMDp------lLEEYFp--------------sLKHQAVSFCKV 219
gi 2105139  206   TPHQIRDVQRSWenir------ndrnALVSSIFVKLFKETp------rIQKFFakfanvavdslagnaeYEKQIALVADR 273
gi 17509143  16   KPEGRKADNQILnsy---------qkSIVRNAWRHMSQKGpsncgstiTRRMMarkstig--dildrstLDYHNLQIVEF  84
gi 17541936 162   DKESCEVVADSWrlvesrssaaetsaCFGLFVFQRVFSKIp------mLRPLFglsesddvfdlpdnhpVRRHARLFTSI 235
gi 17570331  58   SPEHQKLIKRSWnri--------pkaQFGRASLEAFITAAq------vTHAIFvdk-----------etENRHVKYFVDL 112
gi 25155167 144   SPYQQKLLVQCWpniyt----tgasgPFANSLYSTLSSRNa------kAKELLakadgvavfsksdfdcSVMHCRVTVEI 213
gi 28571530  36   TLSERLALRQAWnlvr------pferRYGQDVFYSFLNDYy------wGIKKFrngae------lnvkaLHSHALRFINF  97

                          90       100       110       120       130       140       150
                  ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....
Feature 1                                  ##              #  #                            #
1ASH         73   CHVLCAtyddret-fnaytrELLDRHARdhvh--mppeVWTDFWKLFEEYLgkktt---ldeptKQAWHEIGREFAKEI 145
1HDA_A       66   LTKAVEhlddl----pgalsELSDLHAHklrv---dpvNFKLLSHSLLVTLashlps-dftpavHASLDKFLANVSTVL 136
gi 422406    74   FDNCVKnvgndkg-fqkviaDMSGPHVArpit---hgsYNDLRGVIYDSMHlds--------thGAAWNKMMDNFFYVF 140
gi 1730834  220   LDSAIDnlenvhv-lddyivKLGKRHSRilgi---ktvGFEVMGKAFMTTLqdrfgs-fltlelKNLWGQLYSYLANCM 293
gi 2105139  274   LDTMISamddklq-llgninYMRYTHTErgi----praPWEDFSRLLLDVLgskgvstddldswKGVLAVFVNGVSPIK 347
gi 17509143  85   LQKVMQsldepdk-isklcqEIGQKHAKyrrskgmkidYWDKLGEAITETIreyqgw-kihresLRAATVLVSYVVDQL 161
gi 17541936 236   LHISVKnvdeleaqvaptvfKYGERHYRpditphmteeNVRVFCAQIVCTVfdflrdteatpkcAESWIELMRYLGQKL 314
gi 17570331 113   VQSCVDnlenletgvkpwldLIGRGHANfki----tgkHWEKFGESLLTTAtewngpgrrhketVKAWMVMSSFLADRL 187
gi 25155167 214   LDTVIKnldndharitqyltEIGKQHRHlkaeg-lssaVWDDLGDTIMDCArrrceavrkhkelRRAWLAIIAYIMDNL 291
gi 28571530  98   FGLLIEekdpvv--fqlminDNNHTHNRchv----gsvNIGHLAQALVDYVlkvfhk-vsspslEQGLSKLVEKFQNYQ 169
```

CDD uses RPS-BLAST: reverse position-specific

Purpose: to find conserved domains in the query sequence

Query = your favorite protein

Database = set of many position-specific scoring matrices (PSSMs), i.e. a set of MSAs

CDD is related to PSI-BLAST, but distinct

CDD searches against profiles generated from pre-selected alignments

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

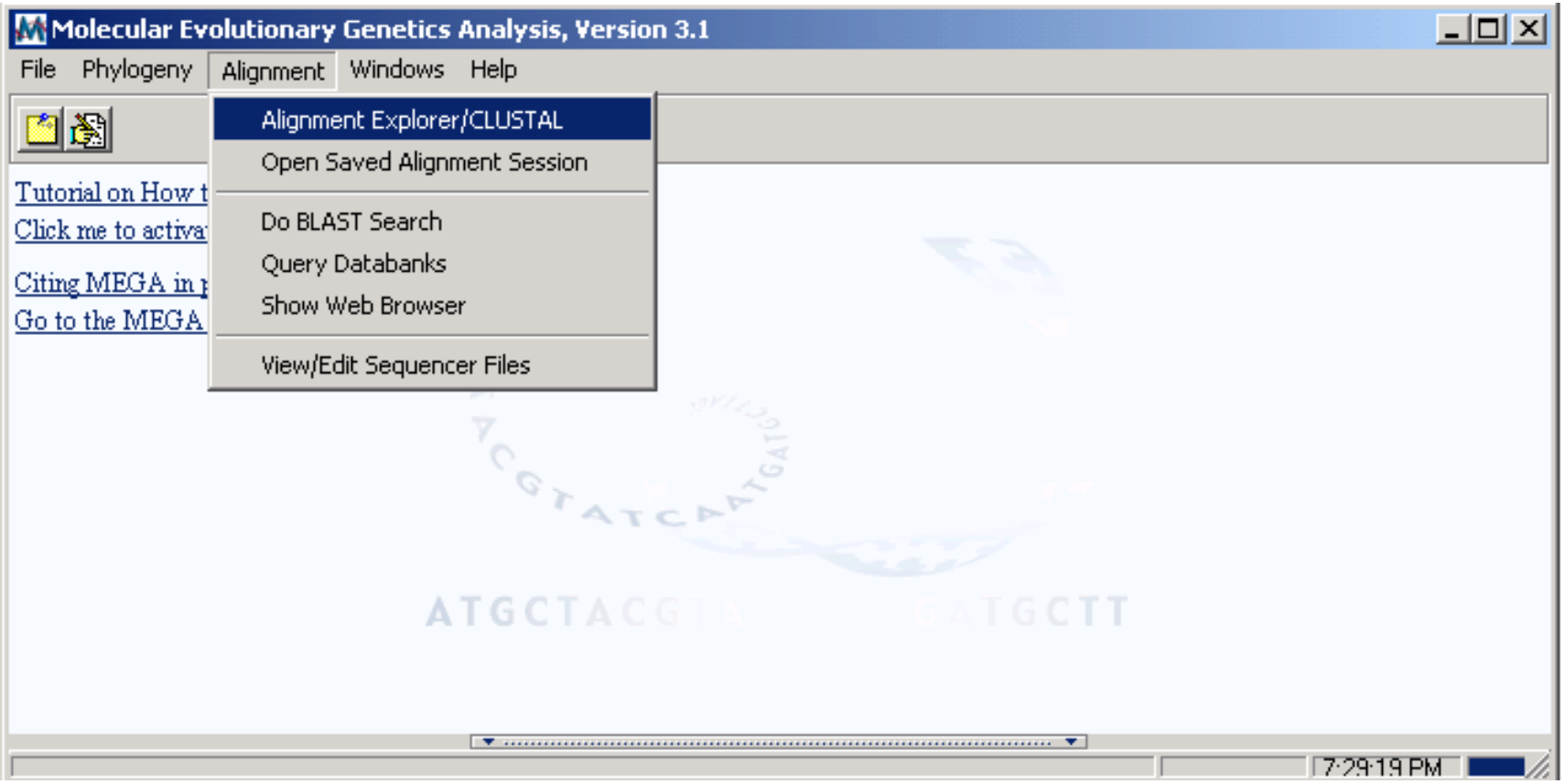
[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

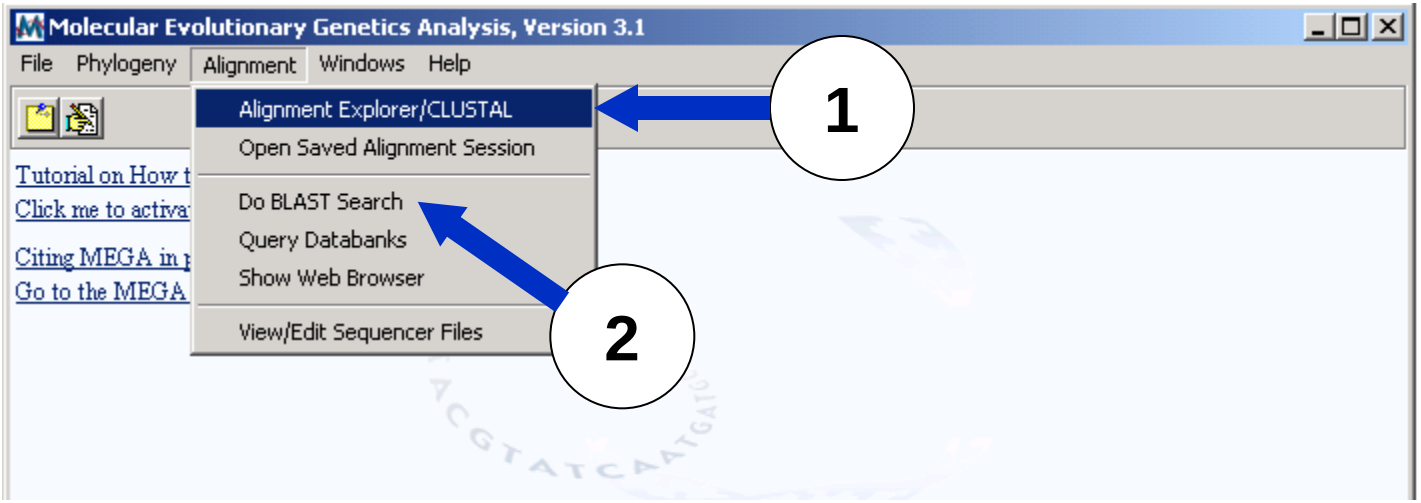
MEGA version 4:
Molecular Evolutionary Genetics Analysis

Download from www.megasoftware.net

MEGA version 4:
Molecular Evolutionary Genetics Analysis



MEGA version 4:
Molecular Evolutionary Genetics Analysis

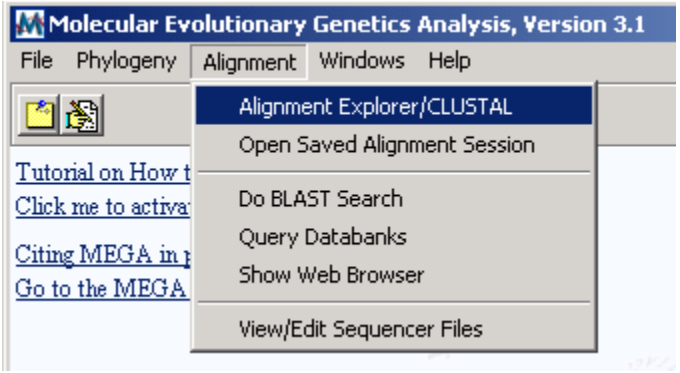


Two ways to create a multiple sequence alignment

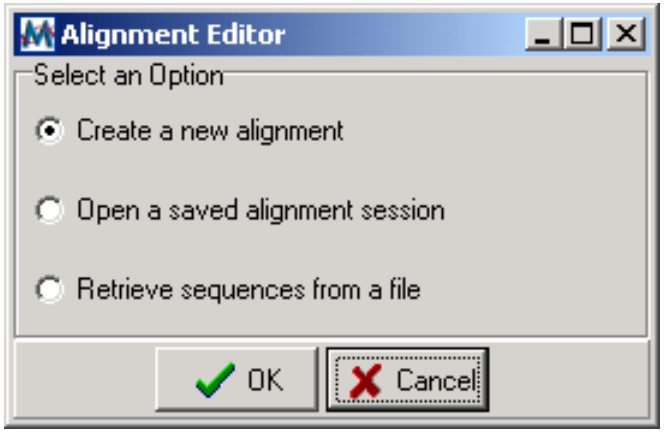
1. Open the Alignment Explorer, paste in a FASTA MSA

2. Select a DNA query, do a BLAST search

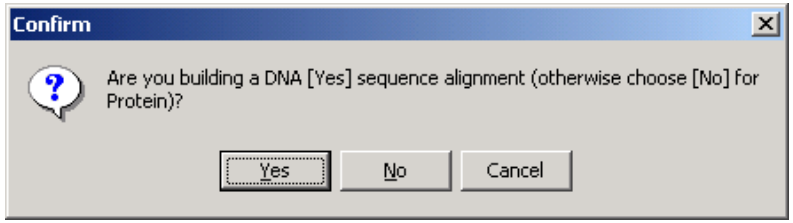
Once your sequences are in MEGA, you can run ClustalW then make trees and do phylogenetic analyses



[1] Open the Alignment Explorer



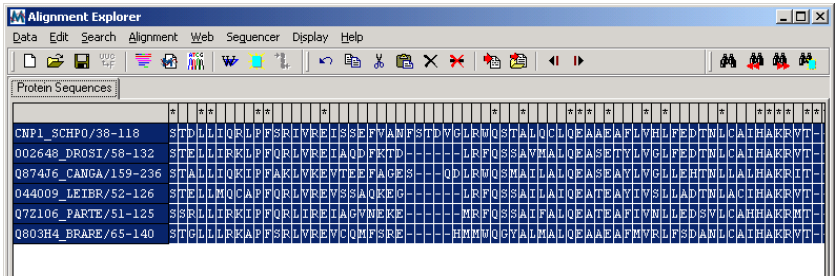
[2] Select “Create a new alignment”



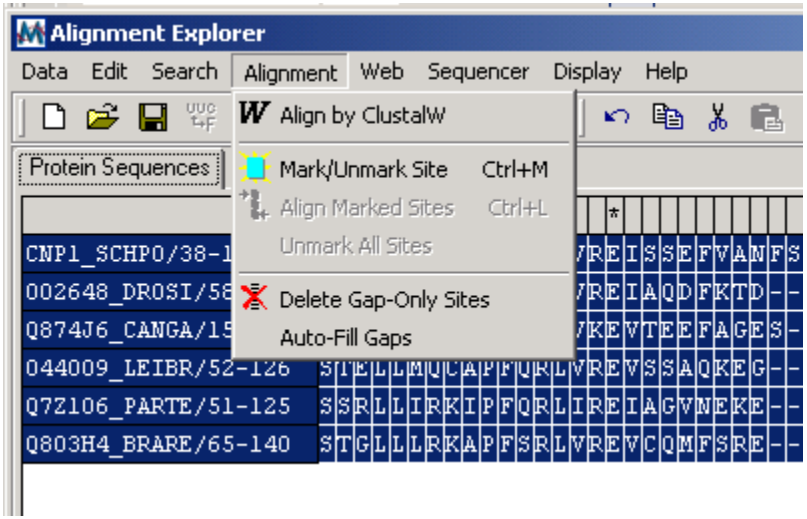
[3] Click yes (for DNA) or no (for protein)

```
>CNP1_SCHPO/38-118
STDLLIQRLPFSRIVREISSEFVANFSTDVGLRWQSTALQCLQEAAEAFLVHLFEDTNLC
AIHAKRVT--IMQRDMQLARRIR
>O02648_DROSI/58-132
STELLIRKLPFQRLVREIAQDFKTD------LRFQSSAVMALQEASETYLVGLFEDTNLC
AIHAKRVT--IMPKDIQLARRIR
>Q874J6_CANGA/159-236
STALLIQKIPFAKLVKEVTEEFAGES---QDLRWQSMAILALQEASEAYLVGLLEHTNLL
ALHAKRIT--IMKKDMQLARRIR
>O44009_LEIBR/52-126
STELLMQCAPFQRLVREVSSAQKEG------LRFQSSAILAIQEATEAYIVSLLADTNLA
CIHAKRVT--IQPKDVQLAMRLR
>Q7Z106_PARTE/51-125
```

[4] Find, select, and copy a multiple sequence alignment (e.g. from Pfam; choose FASTA with dashes for gaps)



[5] Paste it into MEGA



[6] If needed, run ClustalW to align the sequences

[7] Save (Ctrl+S) as .mas then exit and save as .meg

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[3] Hidden Markov models (HMMs), Pfam and CDD

[4] MEGA to make a multiple sequence alignment

[5] Multiple alignment of genomic DNA

Multiple sequence alignment of genomic DNA

There are typically few sequences (up to several dozen), each having up to millions of base pairs. Adding more species improves accuracy.

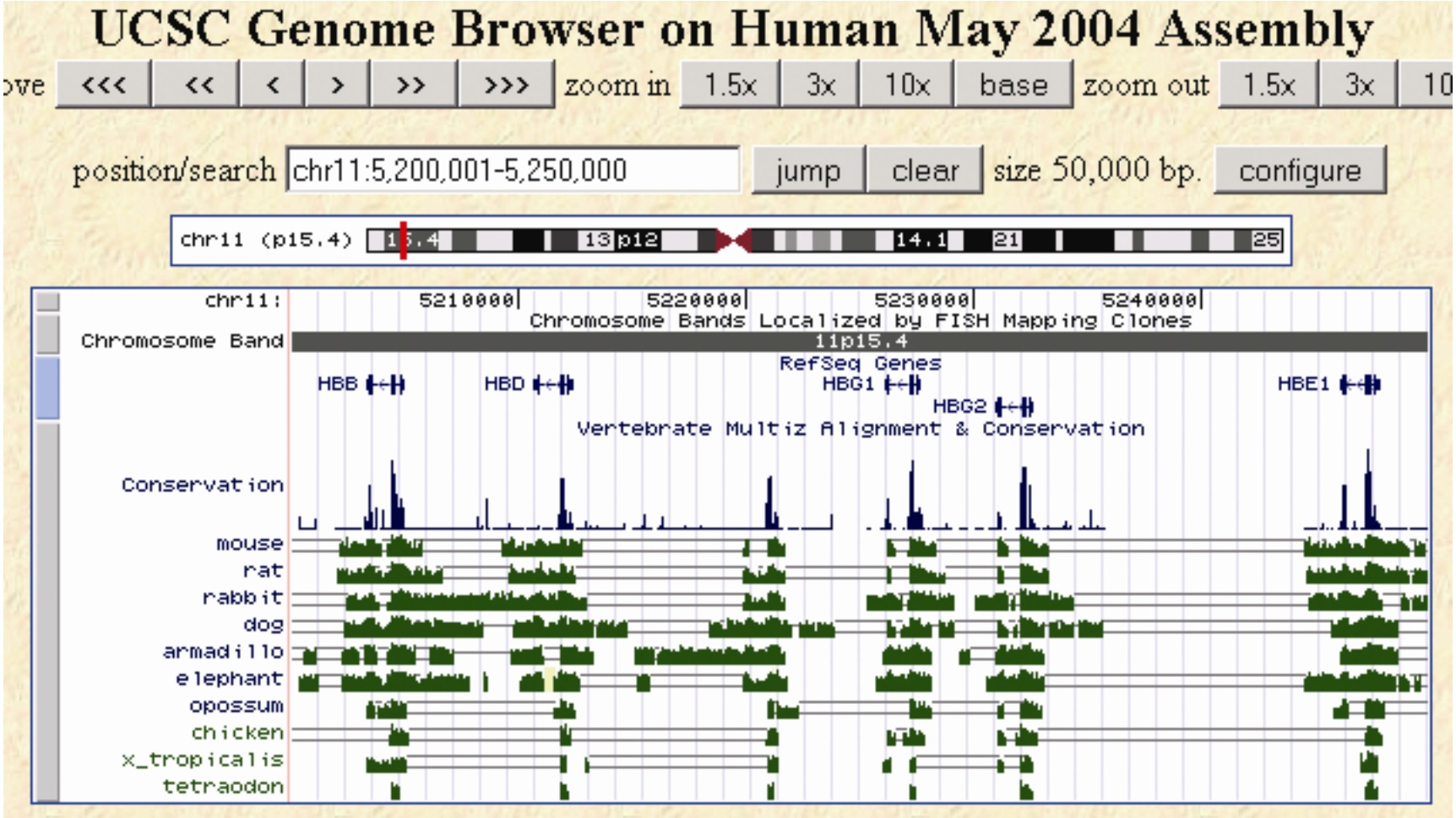
Alignment of divergent sequences often reveals islands of conservation (providing “anchors” for alignment).

Chromosomes are subject to inversions, duplications, deletions, and translocations (often involving millions of base pairs). E.g. human chromosome 2 is derived from the fusion of two acrocentric chromosomes.

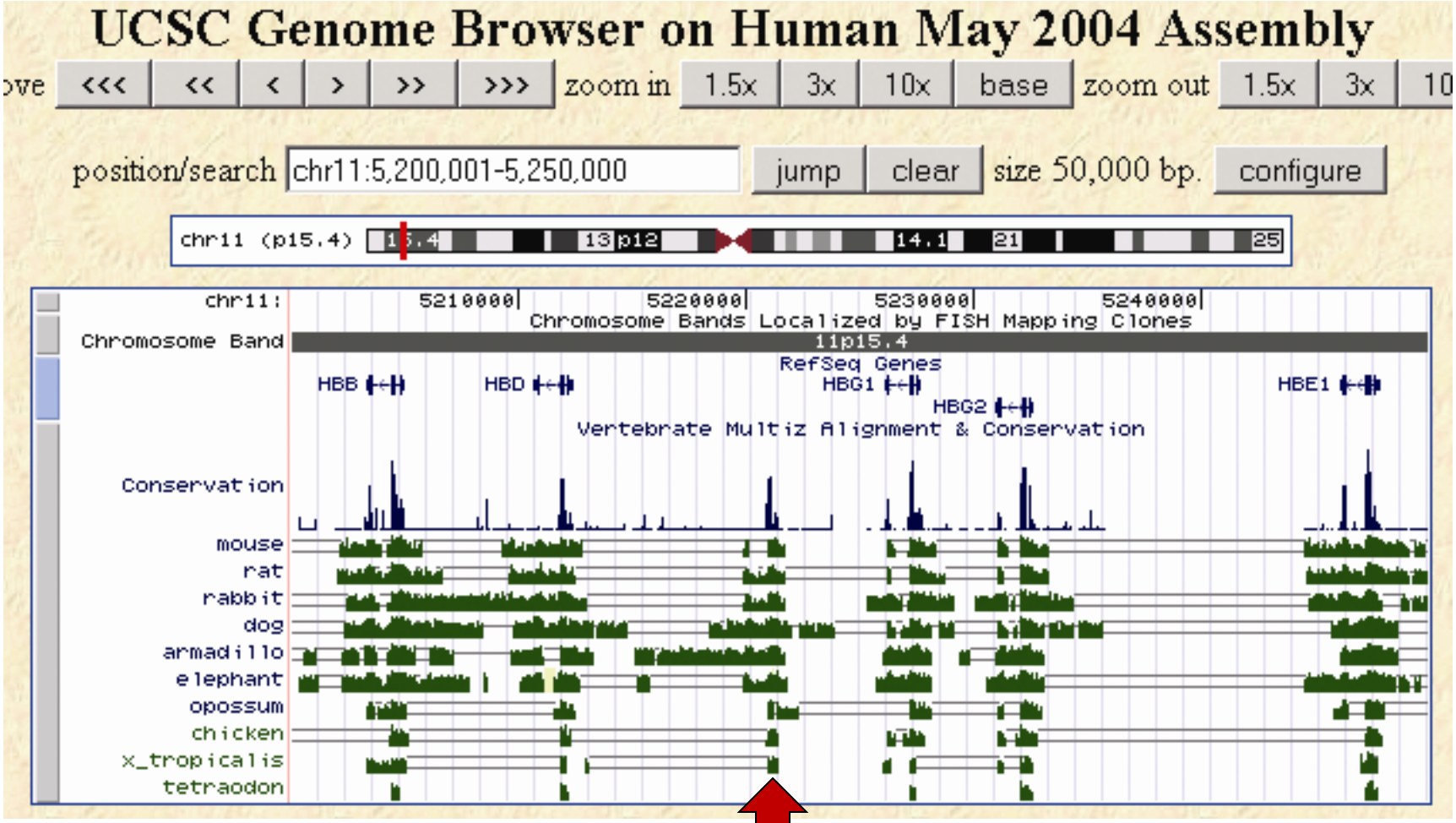
There are no benchmark datasets available.

Multiple alignment of genomic DNA at UCSC

50,000 base pairs (at http://genome.ucsc.edu)

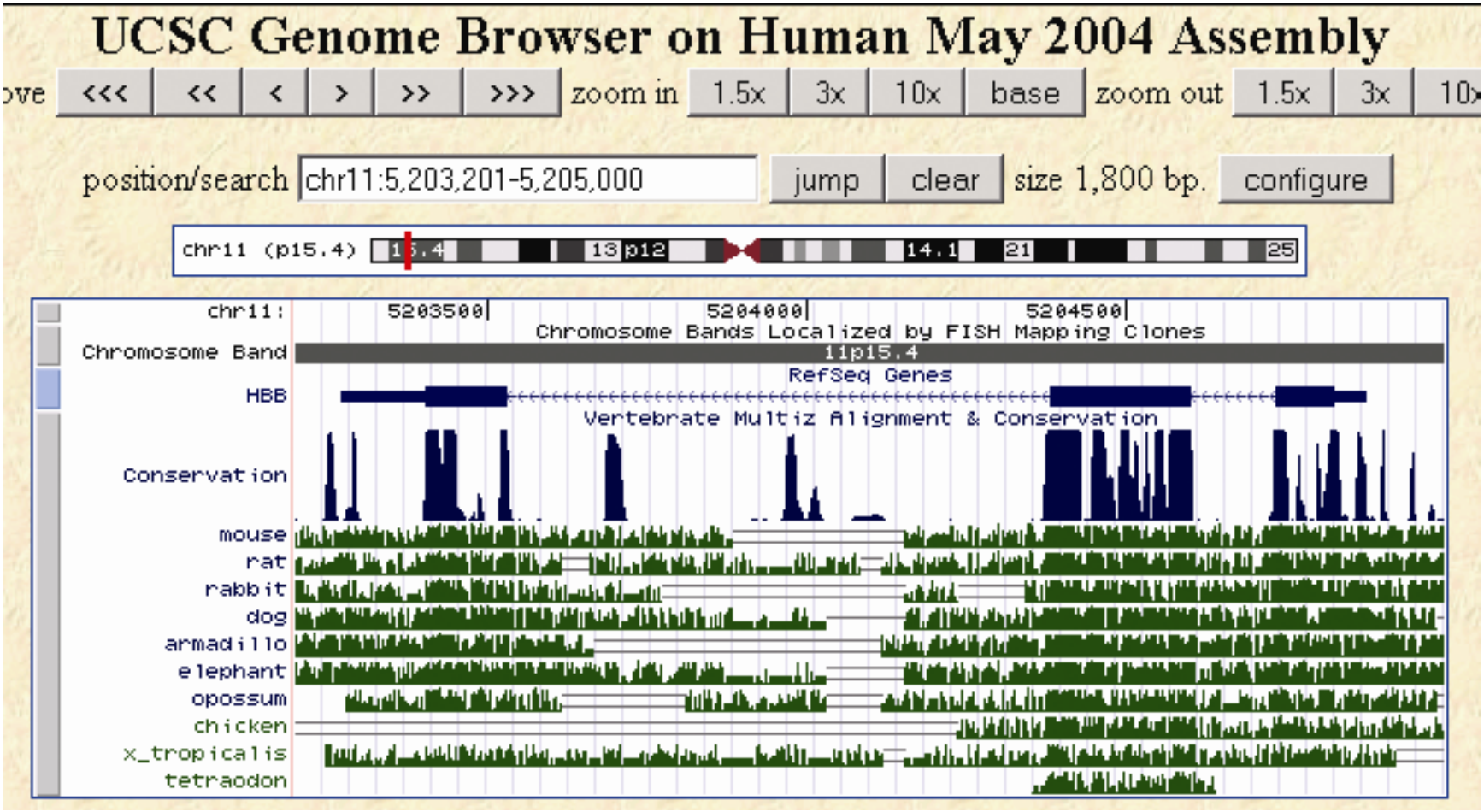


Note conserved regions: exons and regulatory sites (scale: 50,000 base pairs)



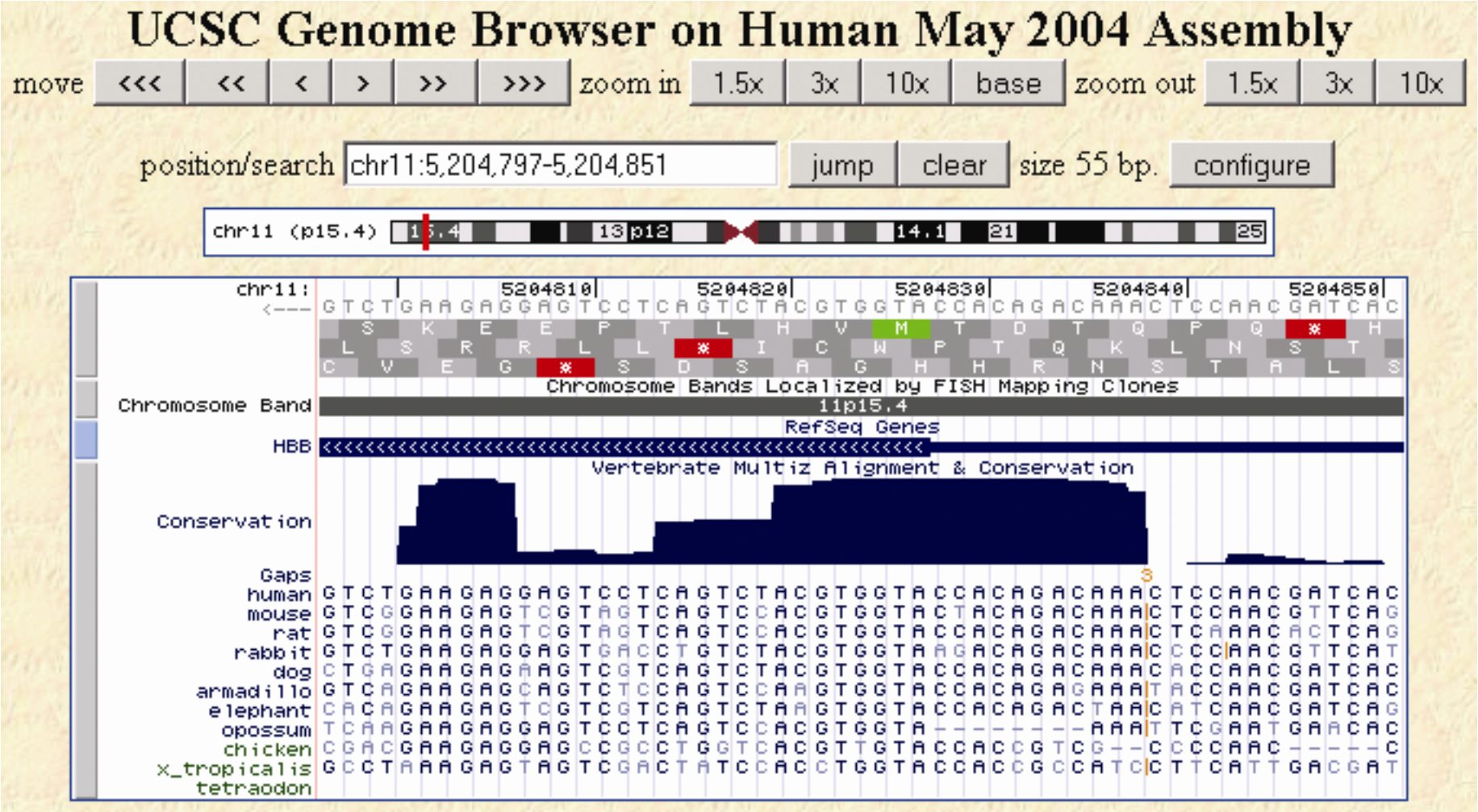
regulatory

Multiple alignment of beta globin gene
scale: 1,800 base pairs



Multiple alignment of beta globin gene

scale: 55 base pairs



This week: please download MEGA software and paste in a set of protein sequences. We'll use MEGA next week to make phylogenetic trees.

Download from www.megasoftware.net